DOCUMENT RESUME

ED 021 737 24 SE 004 635

By- LePage, Wilbur R.; Balabanian, Norman
INTRODUCTION TO ELECTRICAL SCIENCE.
Syracuse Univ., N.Y. Dept. of Electrical Engineering.
Bureau No- BR-5-0796
Pub Date 64
Contract- OEC-4-10-102
Note- 326p.
EDRS Price MF-$1.25 HC-$13.12
Descriptors- *COLLEGE SCIENCE, ELECTRICITY, *ENGINEERING EDUCATION, *INSTRUCTIONAL MATERIALS, PHYSICAL SCIENCES, TEXTBOOKS, UNDERGRADUATE STUDY
Identifiers- United States Office of Education

This text (in mimeographed form) was developed under contract with the United States Office of Education and is intended as material of a first course in the electrical engineering sequence. Introductory concepts such as charge, fields, potential difference, current, and some of the basic physical laws are presented in Chapter I. Subsequent chapters develop concepts of (1) resistive and diode networks, (2) electrostatics, (3) electromagnetism, (4) steady state current analysis, (5) natural response of electric circuits, (6) electric motors, (7) semiconductor theory, (8) transistor amplifiers, and (9) magnetic coupling. (DH)

SE004635
ED021737

INTRODUCTION TO

ELECTRICAL SCIENCE

by

Wilbur R. LePage
and
Norman Balabanian

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Electrical Engineering Department

Syracuse University

Contract No. OE 4-10-102
U.S. Office of Education

Copyright © 1964

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Norman Balabanian

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

INTRODUCTION TO ELECTRICAL ENGINEERING SCIENCE

INDEX

- Ampere, 1-8, 4-28
- Ampere's Law, 4-9
- Armature, 4-26
- Biot-Savant Law, 4-15
- Capacitance, 3-16,17
- Capacitors, 3-25
- Charge, 1-1, 1-22, 3-3
  - test, 1-3
  - point, 3-3
  - stationary test, 3-6
  - bound, 3-24
  - free, 3-24
- Coercivity, 4-25
- Conductance, 1-17
- Conducting materials, 3-14
- Coupling coefficient, 4-44
- Current, 1-8,9
  - branch, 2-10
  - dividers, 2-3
  - loop, 2-10
  - source deactivation, 2-16
  - source equivalent, 2-5
- Diamagnetic material, 4-23
- Dielectric constant, 3-21
  - displacement vector, 3-25
  - breakdown strength, 3-27
- Diode circuits, 2-16
  - ideal, 2-17
  - model, 2-19
- Dipole, 3-5
  - field plot, 3-6
- Electromagnetism, 4-1
- Electromotive force, 4-12
- Electrostatic, 1-6,7, 3-1
- Energy, 1-20
  - storage in elec. field, 3-33
  - storage in magnetic field, 4-54
- Farads, 3-17
- Faraday's law, 4-18
- Ferromagnetic material, 4-20,23
- Field, conservative, 3-14
  - electric, 1-2, 3, 4, 3-1, 4,6,7
  - intensity, 4-22
  - magnetic, 1-7, 4-6
- Flux density, 4-5,15
  - leakage, 4-33
  - linkage theorem, 4-51
  - remnant, 4-25
- Fringing, 4-32
- Gauss' law, 3-8,11,14,16
- Gaussian surface, 3-9,10,11,15,16
- Henrys, 4-40
- Hysteresis
  - dielectric, 3-28
  - loss, 3-34
  - magnetic, 4-25
- Inductance
  - mutual, 4-43
  - self, 4-40
- Isotropic, 4-23,3-29
- Joule, 1-20
- Kirchhoff's laws, 1-10,11,12,13,14, 15,16
- Lenz's law, 4-20
- Load line, 2-20
- Loop equation, 2-8
- Lorentz force, 4-11
- Mmf, 4-28
- Node equations, 2-11
  - datum, 2-11
  - voltages, 2-11
- Norton equivalent, 2-7
- Ohm's law, 1-17, 2-2,3
- Parallel circuits, 2-3
- Paramagnetic material, 4-23
- Permeability, 4-4
- Permeance, 4-38
- Permittivity, 3-4
  - relative, 3-21
- Polarization, 3-24
- Potential difference, 1-3,5, 3-12,14
- Power, 1-20
- Resistance, 1-17
  - forward, 2-17
  - reverse, 2-17
- Resistivity, 1-17,19
- Saturation, 4-21
- Series circuits, 2-1
- Superposition principle, 2-15
- Thevenin's theorem, 2-7
  - equivalent voltage, 2-7
  - equivalent resistance, 2-7
- Transformer, 4-26,44
- Transients
  - R-C circuits, 3-29
  - R-L circuits, 4-47
- Voltage, 1-3,7,20
  - divider, 2-1
  - motional induced, 4-12
  - source deactivation, 2-16
  - source equivalent, 2-5
- Watt, 1-21
- Weber, 4-5

Chapter 1

INTRODUCTORY CONCEPTS

The notion of *electric charge* has grown as a consequence of man's observations of a large number of phenomena, from lightning to the electric shock obtained by touching a metallic object after walking on a thick rug. Electric charge, as a property of electrons and protons, is a major building block of nature. The atomic theory of matter, at least in its simple form, is known to every schoolboy.

It has been known for more than two centuries that objects, which we now describe as being electrically charged, exert forces of attraction or repulsion on each other. The far-reaching consequences of these forces pervade almost all aspects of modern life -- from electrically-powered industrial and home machinery to telephonic communication, to radio and television, to devices for medical diagnosis and treatment, to the guidance and control of space vehicles. A highly satisfactory explanation of these forces has been achieved through inventing the concept of *electric charge*.

Quantitative laws have been discovered which relate the behavior of charged bodies to their configuration, their positions and orientations, and their states of motion. Application of these laws by engineers has led to the development and design of a host of useful accomplishments -- from communication using laser beams to microwave broiling of hamburgers.

1-1. *Coulomb's Law*

The basic amount of charge is that of an electron. However, this is so small compared to charges whose effects are observable in the world of the laboratory that it is not chosen as a unit of measure. In the rationalized MKSC (meter-kilogram-second-coulomb) system of measurements the unit of charge is called the *coulomb*, named after a Frenchman who first gave a quantitative relationship for the force of attraction or repulsion of one electric charge on another. This quantitative relationship, called *Coulomb's law*, was also named after him. It states that the algebraic value of the force *on* a small stationary electric charge Q_1 *due to* another small stationary electric charge Q_2 is proportional to Q_1 and Q_2 and inversely proportional to the square of the distance r between the charges.

$$F_{21} = K \frac{Q_1 Q_2}{r^2} \qquad (1\text{-}1)$$

The constant of proportionality is dependent on the medium in which the charges are located. (We assume that they are located in a homogeneous medium.) In the rationalized MKSC system the constant is chosen so that the force is $1/4\pi\epsilon$ newtons for two identical charges of 1 coulomb each separated by a distance of 1 meter in vacuum. ϵ is called the permittivity of the medium. When the medium is vacuum the symbol is written ϵ_0 and has the value 8.85×10^{-12}.*

Thus, Coulomb's law is written

$$F_{21} = \frac{Q_1 Q_2}{4\pi\epsilon r^2} \qquad (1\text{-}2)$$

This expression represents the algebraic value of the force. But force is a vector quantity and has direction also. The direction of the force exerted on a charge Q_1 by charge Q_2 is along a line directed from Q_2 to Q_1. It is not necessary to specify whether Q_1 and Q_2 are positive or negative; it is only necessary to treat the charges as algebraic quantities. Note the order of the subscripts on F_{21}: the first subscript specifies the charge, Q_2, which is exerting the force; the second subscript specifies the charge, Q_1, on which the force is exerted.

When more than two charges are present in a region the total force on a given charge is the resultant of the individual forces due to all of the other charges. Since each of these individual forces is a vector, the resultant must be composed as a vector sum. We shall designate vectors by putting an arrow above the symbol, a vector force is designated $\vec{F}$.

1-2. Electric Field

If we single out a charge Q_0 and consider the force $\vec{F}$ on this charge due to all other charges Q_1, Q_2, ..., Q_n, we will have an expression

$$\vec{F}_0 = \vec{F}_{10} + \vec{F}_{20} + \ldots + \vec{F}_{n0} \qquad (1\text{-}3)$$

---

* Actually ϵ_0 is related to the speed of light, c, by the relationship $4\pi\epsilon_0 = 10^7/c^2$. Recognition of this relationship was an important factor in identifying electromagnetic radiation (radio waves) as the same thing as light.

where $\vec{F}_{10}$ is the force on Q_0 due to charge Q_1, $\vec{F}_{20}$ is the force on Q_0 due to Q_2, etc. Each of these partial forces is proportional to Q_0, according to Coulomb's law. Hence Q_0 can be factored from each term of the sum and the result can be written

$$\vec{F}_0 = Q_0 \vec{E} \tag{1-4}$$

where $\vec{E}$ is the vector summation of all the terms after Q_0 has been factored. We see that $\vec{E} = \vec{F}_0/Q_0$, the force on charge Q_0 per unit charge. It does not depend on Q_0; by Coulomb's law it depends on all the other charges and on their distances from Q_0, as well as on the permittivity of the medium containing the charges. We call it the *electric field*. The presence of this electric field is not dependent on the presence of Q_0, which is simply a *test charge* placed somewhere to see if there *is* a force on an electric charge at that point. If we were to double the value of the test charge Q_0, the value of the force on it would be doubled, by Coulomb's law, *assuming that this very doubling of the charge did not change the positions of all other charges*. But a doubled force divided by a double charge leads to the same value of $\vec{E}$. If the *location* of Q_0 is changed without changing the locations of any other charge, the value of r to use in Coulomb's law for each charge will change and, hence, the electric field will change. In a region containing charges, then, the electric field varies from point to point.

1-3. Potential Difference and Voltage

Now consider a charge Q_0 placed in an electric field. (This is a short-hand way of saying "placed in a region in which there is a force on an electric charge at any point.") Since there is a force on Q_0, it will tend to move unless there is a restraining force preventing the charge from moving. Suppose there is no other force except that due to the field $\vec{E}$, and the charge moves a certain distance. Work must be done in moving the charge and the amount of this work equals the distance moved times the component of force along the line of motion. Consider the diagram in Figure 1-1. Suppose Q_0 moves a distance ℓ from point P_1 to P_2 in a straight line.

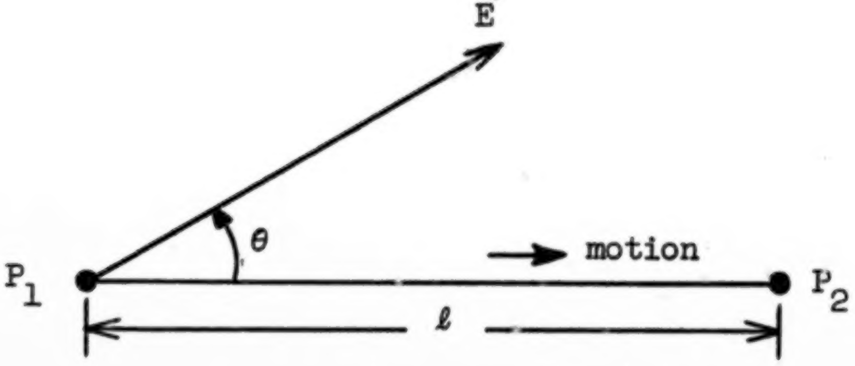


Figure 1-1.

We might be tempted to say the work done is ℓ times Q_0E (the magnitude of the vector) times cos θ, where $Q_0E \cos\theta$ is the component of the force on Q_0 along the line of motion. But since $\vec{E}$ may not have the same direction from point to point as the charge moves, its component ($E \cos\theta$) along the line of motion might change from point to point. To overcome the difficulty, we must find the differential work done in moving along a small displacement, $d\ell$, assuming that the direction of $\vec{E}$ does not change over this small distance. Then, we must integrate to find the total work. Thus,

$$\begin{array}{l}\text{Work done by the electric}\\ \text{field in moving a charge}\\ \text{from } P_1 \text{ to } P_2\end{array} = \int_{P_1}^{P_2} Q_0E \cos\theta \, d\ell \qquad (1\text{-}5)$$

where E and cos θ can change from point to point along the path.

Now, the potential energy that is available to do work by virtue of the fact that Q_0 is in an electric field has been decreased. (This is analogous to a body falling down a hill. The gravitational force exerted on the body when it is at the top of the hill is doing work as it falls. But all the while, the potential energy which was available is decreasing until, when the body reaches the bottom of the hill, it cannot fall any more and its potential energy is reduced to zero.) Returning to the integral in Eq. (1-5), we see that the work done represents the decrease in potential energy as Q_0 moves from P_1 to P_2. If we divide by Q_0, we find

$$\begin{array}{l}\text{Decrease in potential energy per unit}\\ \text{charge in moving a charge from } P_1 \text{ to } P_2\end{array} = \int_{P_1}^{P_2} E \cos\theta \, d\ell \qquad (1\text{-}6)$$

This quantity is extremely important. In order to avoid having to say all those words on the left side every time we want to refer to it, it is given a name; it is called the potential difference. More specifically it is the potential decrease from P_1 to P_2.

Normally it is a difficult job to calculate the magnitude and direction of the electric field at all points along a path spanning two points between which it is desired to know the potential difference. Fortunately, in much of the study of electrical engineering it is not necessary to do this since other relationships can be used besides this integral.

We shall use the term voltage to stand for the potential decrease as calculated from the above integral. By definition, then, the voltage across two points P_1 and P_2 is the decrease in potential energy per unit charge when a charge moves from P_1 to P_2.

Although the preceding discussion was carried out in terms of the electric field as the agency which exerted force and did work, the resulting definition of voltage is more general; the work can be done by the field or by an external agency, like gravity. Hence, the decrease in potential energy may turn out **not to be an actual decrease at** all, but an increase. Equation (1-6) takes this into account, as the resultant voltage is an algebraic quantity and may vary in sign. That is, when the potential energy per unit charge is actually increased, the decrease in potential energy is negative.

We shall use the symbol, v, to stand for voltage. Since the definition involves two points there must be some way of designating from which point to which point the integration in Eq. (1-6) is to be carried out. In Figure 1-2 two points a and b are shown, the voltage between which is under discussion.

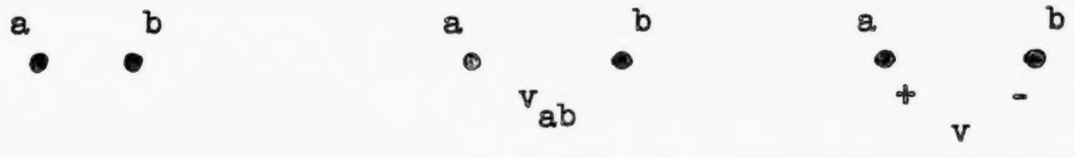


Figure 1-2.

If we go from a to b in the definition of voltage, we will label it v_{ab}. Thus, v_{ab} is the decrease in potential energy per unit charge when a charge moves *from* a *to* b. In a particular situation this number may turn out to be positive or it may be negative. If the number is positive, it means that there is *actually* a decrease in potential energy in going *from* a *to* b. If negative, this means that there is actually an *increase* in potential energy in moving *from* a *to* b. In either case we get the required information about the state of potential energy.

Another way of designating *from* which point *to* which point in the integration of Eq. (1-6) is to use some marking, as in the third part of Figure 1-2. A + sign is placed at the point *from* which we integrate and a - sign at the point *to* which we integrate. Thus, if we write v and put a plus sign at *a* and a minus sign at *b* we mean v_{ab}. (Of course, only the plus sign is enough; once you know which point carries the plus sign, then you know the other point carries the minus sign.)

In discussing Coulomb's law we pointed out that the relationship holds for stationary charges. (The term *electrostatics* is used to designate this branch of the subject.) However, to determine the voltage we talked about moving a charge from one point to another. How can these two notions be reconciled? Actually, in moving a charge about in a field we must think of motions carried out with infinitesimally *small* velocities, otherwise kinetic energy will have to be considered and we no longer have an electrostatic situation. But just what is an infinitesimally small velocity? This is really hard to answer. By this phrase we shall mean, so small that the kinetic energy can be neglected compared with the potential energy involved. To imply that there is some non-zero velocity but that the results based on the static case still apply, the term *quasi-static* conditions is usually used. In everything we consider here we often use the term *static* although we shall be dealing with quasi-static conditions.

In discussing the moving of a charge from one point to another, no consideration was given to the path of motion. Thus, in Figure 1-3 a charge can move from

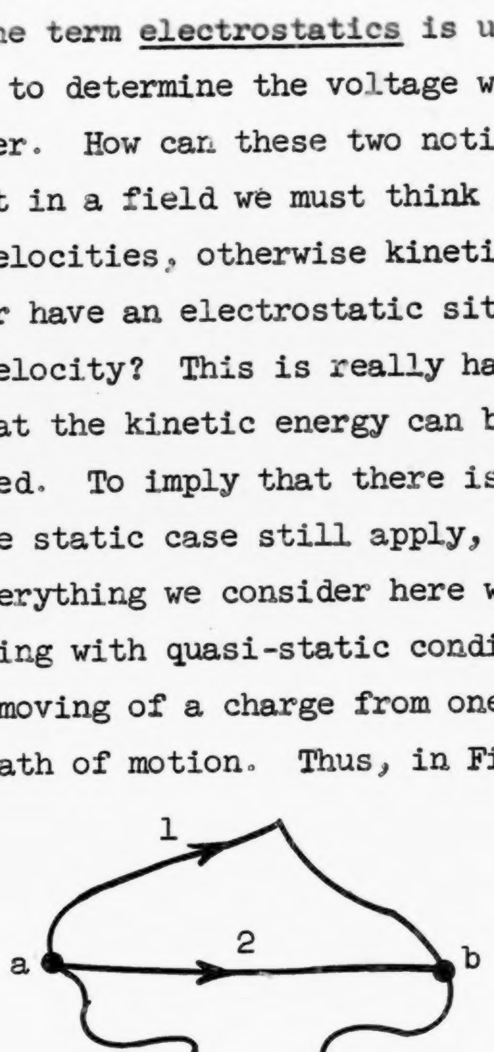


Figure 1-3.

a to b along either of the paths shown, or along any number of other paths. If the notion of voltage is to have a useful meaning, either one of two things must be true about the path taken in going from a to b; either the result must be independent of the path taken -- that is, it will be the same no matter which path is taken -- or a particular path is specified and we agree always to take *that* path.

Although we shall not do so here, it is an easy matter to show that for the *electrostatic* case (charges stationary), the work done in moving a charge between two points is independent of the path taken. The same thing is true even in the case when charges are moving, provided that the resulting currents are constant with time -- so-called *direct current* (dc). In these cases the term "voltage" will have an unambiguous meaning. (Note that in the case of gravitational force a similar result is obtained. That is, if we take a stone falling down a hill, its decrease in potential energy after falling a certain vertical distance is the same no matter how it bounded around in falling that distance.)

But we are also interested in situations where the currents *do* vary with time. In such cases the result of the integration in Eq. (1-6) *will depend* on the path taken. But even in these cases we can make the term, "voltage", meaningful by agreeing as to the path taken. From previous studies you are no doubt familiar with the fact that electric currents are accompanied by *magnetic fields*. (Further discussion of this subject will take place in Chapter 4.) If the currents are time-varying, then so also will be the resulting magnetic field. In order to deal with such a situation, we assume that all time-varying magnetic fields are *localized*; that is, they are *lumped* or *concentrated* in one or more devices, as shown in Figure 1-4, and the effects produced are accounted for at the terminals of the lumped device.

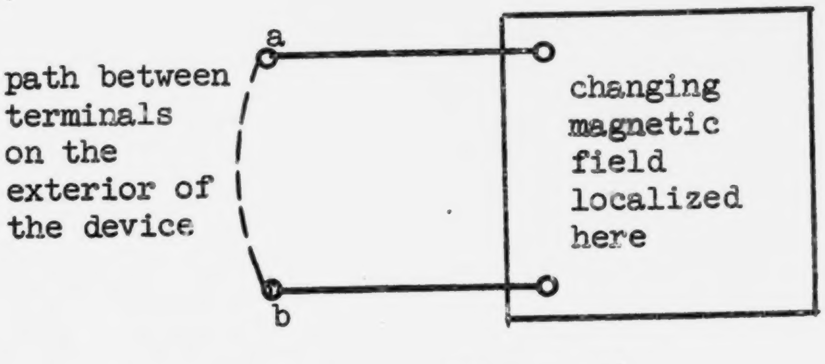


Figure 1-4.

We agree never to choose a path between the terminals a-b that goes inside the device, but always to take a path outside the device. With this agreement the notion of voltage between a and b will again be meaningful.

1-4. Current

If electric charges were all stationary, they would be of very little interest to anyone. They give rise to important observable electric phenomena when they are in motion. Electric charge in motion constitutes a current. Before we define current more carefully, note that both negative and positive charge can be moving, and that each can be moving in any direction.

In developing a definition of current, suppose we consider charges moving along a discrete path, like a conducting wire. Suppose further that only positive charges are moving and we count the amount of charge passing a particular cross section of the wire in a given time interval, say half an hour. We can then say how much charge came by in the half hour, but we won't know whether it came across faster over part of the time and slower over the rest of the time, or whether the rate at which the charge passed the cross section was the same for the whole interval. To remedy this situation we can make the observation interval smaller and smaller. Suppose we take the observation interval to be 2 seconds. We can add up the amount of charge coming by in each 2-second interval. The average rate of passing of the charge equals the amount per 2-second interval divided by 2. But still we can't say anything about the rate of flow within each interval, whether it speeded up or slowed down as the interval of time progressed from 0 to 2 seconds. Suppose we make the observation interval still smaller, say Δt, and we find the amount of charge passing a cross section during this interval to be Δq. The average rate over the time Δt at which charge is moving past a given cross section is then $\Delta q/\Delta t$. In the limit as we let the observation interval get smaller and smaller, we define the electric current i to be

$$i = \frac{dq}{dt} = \lim_{\Delta t \to 0} \frac{\Delta q}{\Delta t} \qquad (1\text{-}7)$$

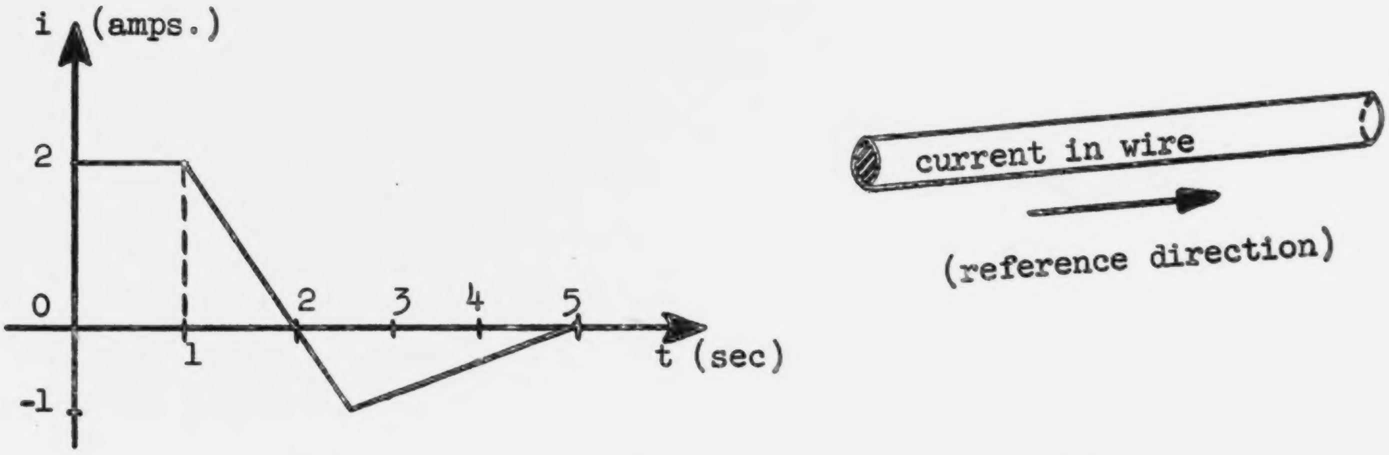
Current has the dimensions of coulombs per second and its unit is the ampere.

But, "moving past a cross-sectional area" can be in either one direction or the other. So we arbitrarily choose a particular direction as our reference. If positive charges go past a cross section in that direction we say the current is positive; if they go in the opposite direction we say the current is negative.

But how about negative charges? Suppose a positive and a negative charge of equal magnitude move at the same rate in the same direction. The net flow of charge will be zero. This means that a negative charge moving in one direction is equivalent to a positive charge of equal magnitude moving in the opposite direction. It is, therefore, unimportant to know whether the current is caused by the motion of positive charge or negative charge or both.

To summarize: when charges are flowing in a discrete path, we arbitrarily assign a direction to be the reference direction. We designate the current to be positive over a time interval if during this time positive charges are moving in the reference direction or negative charges are moving opposite to the reference direction. The reference direction is indicated by an arrow drawn beside the path of charge flow.

In Figure 1-5 is shown a graph of current carried by a wire as a function of time. The reference direction of the current is shown as being to the right.



(a) Figure 1-5. (b)

It is required to answer the following questions:

1. In what direction are positive or negative charges moving in the wire over the interval from 0 to 5 seconds?

2. How much charge and of what sign has passed a cross section of the wire toward the right after 1, 2 and 5 seconds?

(Try to answer these questions before reading on.)

1. The reference direction is to the right. When the current is positive, positive charge is actually moving to the right. When the current is negative, positive charge is actually moving to the left. Thus, from 0 to 2 seconds positive charge is actually moving to the right and from 2 to 5 seconds positive charge is actually moving to the left. Since negative charge moving in one direction is equivalent to an equal positive charge moving in the other direction, the same result in each of the intervals can be described in terms of a negative charge actually moving in a direction opposite to the direction in which the positive charge is moving.

2. Since current is the time derivative of charge transferred, then the amount of charge must be equal to the integral of current.

$$q = \int_0^t i \, dt \qquad (1\text{-}8)$$

An integral can be interpreted as the area under a curve. Hence, after 1 second the area under the curve (Figure 1-5) is 2 times 1 ampere-seconds = 2 coulombs. It is positive, so a net positive charge has passed to the right, or a net negative charge has passed to the left. At 2 seconds, the current is zero but the area under the curve from 0 to 2 is not; it is 2 + 1 = 3 coulombs. (Positive charge to the right or negative charge to the left.) From 2 to 5 seconds the current is negative, so the net area under the curve is being reduced; it is 3 - 3 x 1/2 = 3/2 coulombs.

In order to avoid having to say "positive charge to the left, negative charge to the right" in describing a current, we will henceforth assume that only positive charges are involved. So, when a current is positive, we will say that charge (meaning positive charge) is actually flowing in the reference direction; and when a current is negative, charge (meaning positive charge) is actually flowing opposite to the reference direction.

1-5. Kirchhoff's Laws

(Read the programmed booklet titled "Kirchhoff's Laws" for further instruction in the subject of this section.)

Kirchhoff's current law is a statement expressing the fact that electric charge cannot accumulate at a node or be generated there. Figure 1-6 shows four

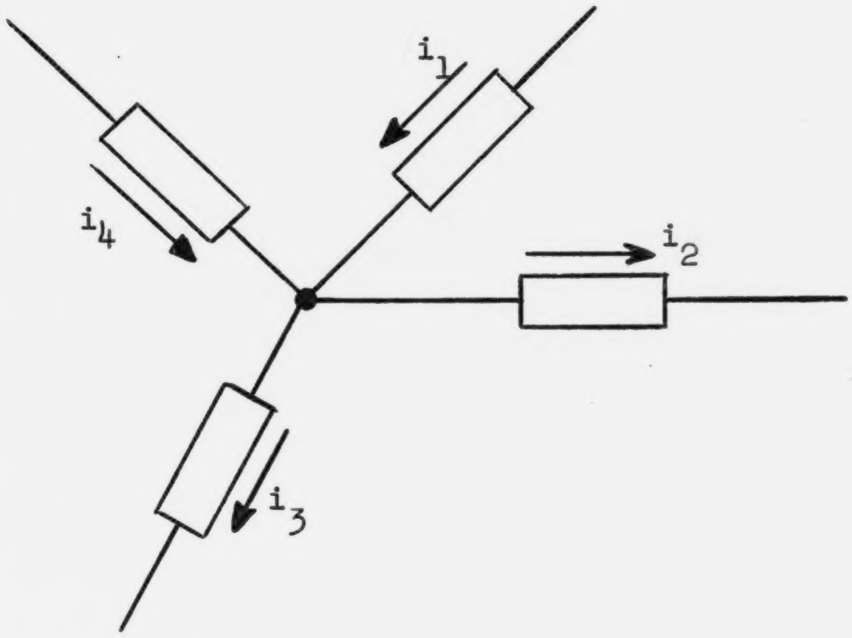


Figure 1-6.

branches joined at a node. The branch currents are each arbitrarily assigned a reference direction shown by an arrow alongside the branch. Kirchhoff's current law can be stated in the following forms:

1. The algebraic sum of all current leaving any node (or junction) is zero at each instant of time (for the example, $-i_1 + i_2 + i_3 - i_4 = 0$); or

2. the algebraic sum of all currents entering any node is zero at each instant of time (for the example, $i_1 - i_2 - i_3 + i_4 = 0$); or

3. the sum of currents with references directed toward a node is equal at each instant of time to the sum of currents with references directed away from the node (for the example, $i_1 + i_4 = i_2 + i_3$).

In the first form, the node reference is 'leaving' while in the second form the node reference is 'entering'. In either case, if a branch reference coincides with the node reference, the corresponding term will carry a negative sign in the mathematical expression; if a branch reference is opposite to the node reference, the corresponding term will carry a negative sign.

Figure 1-7 shows a network having 4 nodes. (For simplicity the rectangles are omitted but every line shown represents a branch, not just a connection.) A reference direction for each branch current is arbitrarily

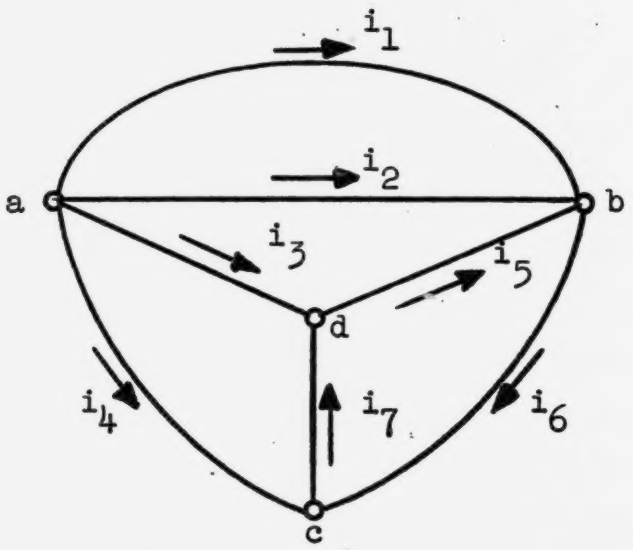


Figure 1-7.

picked. It is desired to apply Kcl and write a set of equations, one for each node. We shall do this, taking 'leaving' as the reference for each node. The result is the following. (You should write your own set of equations and check against these. It doesn't matter if your terms are not in the same order.)

$$\begin{array}{lrrrrrrrl}
\text{For node a:} & i_1 & + i_2 & + i_3 & + i_4 & & & & = 0 \\
\text{For node b:} & -i_1 & - i_2 & & & -i_5 & + i_6 & & = 0 \\
\text{For node c:} & & & & - i_4 & & - i_6 & + i_7 & = 0 \\
\text{For node d:} & & & - i_3 & & +i_5 & & - i_7 & = 0
\end{array} \qquad (1\text{-}9)$$

Additional Comments*

The objective in this problem was simply to give some practice in applying Kirchhoff's current law. However, having written the equations, it is possible to examine them to see if any additional information can be obtained. Notice how the equations have been written, with the terms for a given current appearing in a vertical column. Something very curious can be detected: each current appears twice in a column, once with a plus sign and once with a minus sign. Hence, if the equations are all added, the result will be identically zero! That is,

$$\text{Eq. a} + \text{Eq. b} + \text{Eq. c} + \text{Eq. d} = 0 \qquad (1\text{-}10)$$

---

*Subsections titled Additional Comment are for the purpose of introducing those who are interested to topics beyond the scope of the material for this course. No one is required to read these sections, but they will help any who do reach a deeper understanding.

From this it follows by solving for Eq. d that

$$\text{Eq. d} = -(\text{Eq. a} + \text{Eq. b} + \text{Eq. c}) \qquad (1\text{-}11)$$

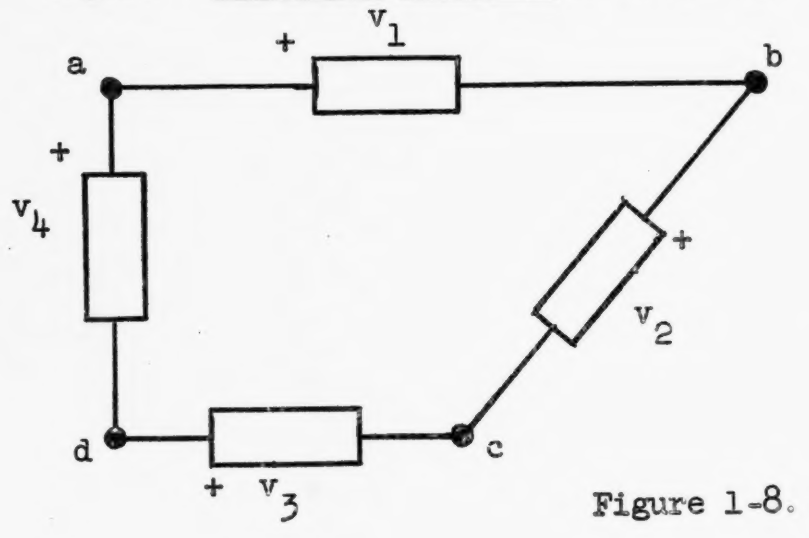
Actually, instead of Eq. d, we can solve for any one of the others and discover that

any one of the four equations = negative sum of all the others

which means that they are not all independent; if all but one are known, this one follows as the negative sum of the others. Verify this by adding the last three in Eq. (1-9) and comparing this sum with the first equation.

This result, which was found to be true for this example, is actually quite general and can be easily demonstrated. That is, for any network having N_n nodes, only N_n-1 independent equations expressing Kirchhoff's current law can be written. There will be more about this in the next chapter.

We turn next to Kirchhoff's voltage law. Figure 1-8 shows four branches forming a closed path. The branch voltages have each been arbitrarily assigned a reference polarity as shown by the plus sign.



$v_1 = v_{ab}$

$v_2 = v_{bc}$

$v_3 = v_{dc} = -v_{cd}$

$v_4 = v_{ad} = -v_{da}$

Figure 1-8.

Kirchhoff's voltage law states that:

1. The algebraic sum of all voltages around any closed path in an electric network (traversed either clockwise or counterclockwise) is zero at each instant of time. (In the example, going clockwise, $v_1 + v_2 - v_3 - v_4 = 0$); or

2. around any closed path and at each instant of time, the sum of voltages with clockwise references is equal to the sum of voltages with counterclockwise references. (In the example, $v_1 + v_2 = v_3 + v_4$.)

In the first form, if a branch reference coincides with the loop reference (that is, the plus sign is encountered first when traversing the branch in the loop reference direction, which may be either clockwise or counterclockwise), the corresponding term will carry a positive sign in the mathematical expression; otherwise, a negative sign.

Kirchhoff's voltage law can be taken as a basic postulate. But if we consider the electrostatic case only, Kvl follows from the definition of voltage. This is easy to appreciate if we remember that in this case the 'voltage' is independent of the path. That is, in Figure 1-9 we can go from a to b along either the upper path or the lower path and the voltage (the decrease in potential energy per unit charge) will be the same: $v_1 = v_2$ or $v_1 - v_2 = 0$, which is Kvl for Figure 1-9.

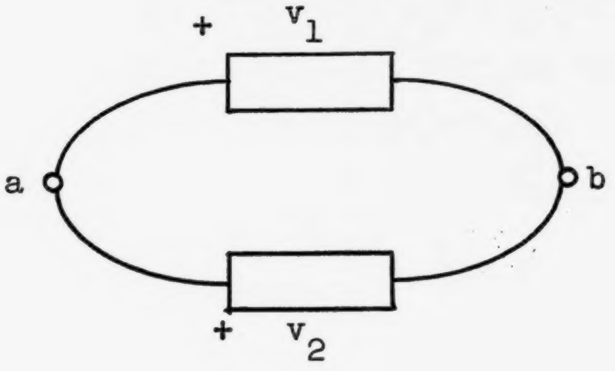


Figure 1-9.

For the general case of lumped networks we agree that in discussing voltage we will always take external paths between the terminals of the branches. As long as we do, it doesn't matter what combination of branches we traverse, the voltage between two points will be the same. Thus, in Figure 1-8 the voltage from a to b must be the same whether we go directly from a to b or go from a to d to c to b. Thus, $v_1 = v_4 + v_3 - v_2$, which can be written $v_1 + v_2 = v_3 + v_4$.

Note that the assumption of a lumped network means that there is no changing magnetic flux passing through the closed path in Figure 1-8.

The diagram in Figure 1-10 shows a network having three closed paths, or loops, as shown by the dashed arrows. (To avoid confusion, the diagram is redrawn without the arrows.) A reference polarity for each branch voltage is arbitrarily

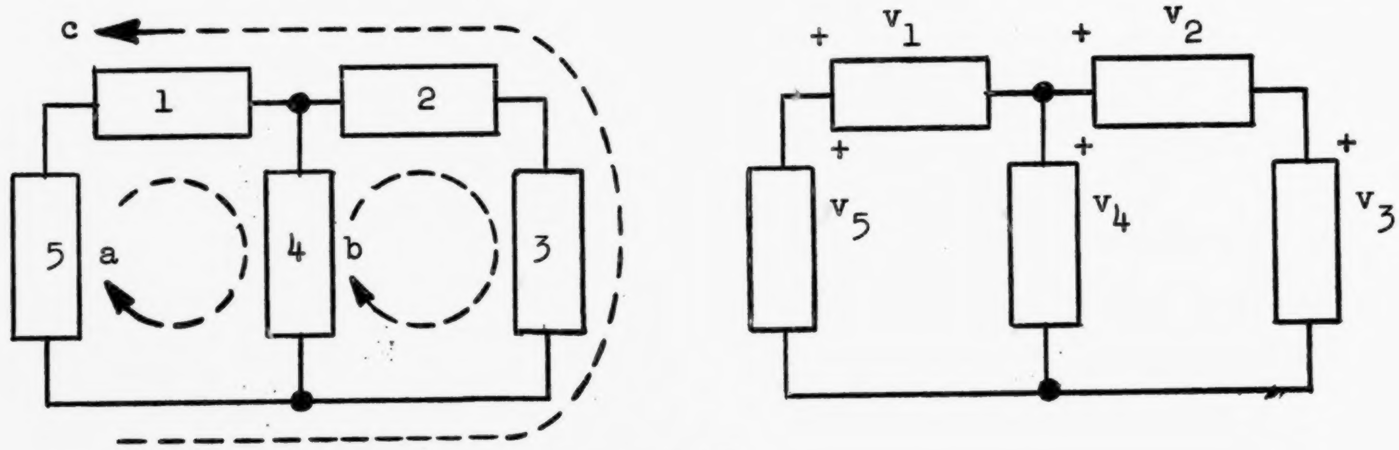


Figure 1-10

assigned. It is desired to apply Kirchhoff's voltage law to write a set of equations, one for each loop. (You should write your own set of equations before going on and check them against the ones below. Don't worry about the terms in your equations being in a different order from these.) Here is the result.

$$\begin{aligned}
&\text{loop a:} \quad v_1 \qquad\qquad\qquad + v_4 - v_5 = 0\\
&\text{loop b:} \qquad\quad v_2 + v_3 - v_4 \qquad\; = 0 \qquad\qquad (1\text{-}12)\\
&\text{loop c:} \quad -v_1 - v_2 - v_3 \qquad + v_5 = 0
\end{aligned}$$

Additional Comments

Let us again examine these equations for any additional information we can gather from them. Note again that each voltage appears twice in a vertical column, once with a plus sign and once with a minus sign. If the equations are all added, the result will, therefore, be identically zero. From this it again follows that any one of the equations can be obtained, once the other two are known. The third equation, for example, the one around the outside contour of the network, is just the negative sum of the other two. (Clearly, if this equation had been written in a clockwise sense instead of the other way, it would have been obtained as the positive sum of the other two equations.)

This result, unlike the corresponding one for Kirchhoff's current law, is not general. In more complicated networks, there are many more closed paths than there seem to be. For example, in Figure 1-11 only one additional branch has been added to the previous network. In addition to the previous

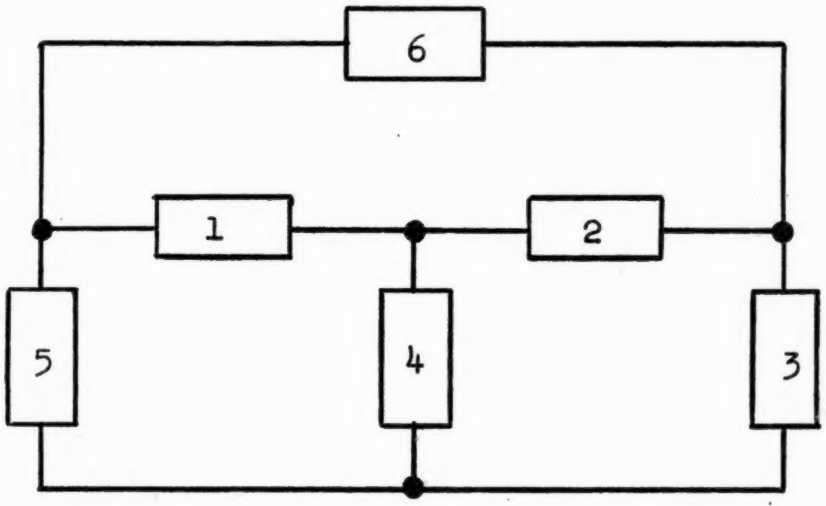


Figure 1-11

3 closed paths, there are now 4 more. Indicating these paths by listing the branches lying on them, these closed paths are 3-5-6, 1-2-6, 1-4-3-6 and 2-6-5-4. Hence, there are a total of 7 closed paths to which Kvl can be applied.

For a given network, it is easy to count the number of junctions to find how many total Kirchhoff current law equations can be written. Of these, all but one are independent. But the situation is different for the number of closed paths. In fact, there is no way of telling from the number of nodes or branches the total number of closed paths a network will have, short of actually finding them all. But fortunately we are not interested in the total number of Kvl equations in a network, only in the independent ones; and these it is possible to tell. If a network has N_n nodes and N_b branches, then there are $(N_b - N_n + 1)$ independent Kvl equations. (We shall not prove this result here.) In Figure 1-11, for example, there are 6 branches and 4 nodes; hence, there should be $N_b - N_n + 1 = 6 - 4 + 1 = 3$ independent Kvl equations. Verify this relationship also for Figure 1-10.

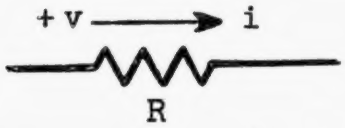
Write Kvl equations around loop 1-4-5, 2-3-4 and 1-2-6 in Figure 1-11 and notice that no one can be obtained from the other two, showing that all three are independent. Then write a Kvl equation for any one of the other closed paths and then try to obtain it by certain combinations of the first three equations.

1-6. Ohm's Law and Resistance

(Read the programmed text booklet titled "Ohm's Law and Sources" for further instruction in the subject of this section.)

By empirically observing the relationship between the voltage and current in metals, it is found that the current is almost directly proportional to voltage. On this basis we introduce the notion of a hypothetical device called an ideal resistor whose voltage and current are exactly proportional. Then, Ohm's law is

$$v = Ri \qquad (1\text{-}13)$$



where R is a constant called the resistance whose unit is the ohm. This relationship applies for the selection of voltage and current references shown. If either of these is reversed, the equation will become $v = -Ri$.

The reciprocal of resistance is conductance G, measured in mhos. Thus, Ohm's law can also be written as

$$i = Gv \qquad (1\text{-}14)$$

Physical resistors (the actual physical devices as distinct from the ideal models) have properties that diverge more or less from the ideal. Although other materials, like carbon, are used in the manufacture of resistors, most resistors are made of metallic wire. In considering the possible factors on which the resistance of a metallic resistor depends, we would no doubt expect the physical properties of the material -- that is, how good a conductor it is -- to have an influence. Other things being equal, we would expect the resistance to be different if one were made of copper or made of aluminum or steel. And for the same material, we would certainly expect the geometry or the dimensions of the conductor to be important. Well, it is possible to derive an expression for the resistance of a piece of metal by using the atomic model for metals and making assumptions on the manner in which the electrons move about under the influence of an electric field and the manner in which the resulting current is distributed within the metal. This expression is:

$$R = \rho \frac{\ell}{A} \qquad (1\text{-}15)$$

where ℓ is the length of the wire and A is its cross sectional area. The quantity ρ is called the resistivity and is a property of the material. (From the equation, you can determine that its dimensions are ________________.)

The resistivity of a material depends on such things as the mass and charge of an electron, the density of free electrons in the material, the average velocity with which they move, and the average distance an electron moves before colliding with another particle. For our purposes, it is enough to know that there is considerable variation of resistivity among materials and that resistivities can be determined by measurement. Table 1-1 shows the resistivities of a number of materials.

Any condition that influences one or more of the quantities on which the resistivity depends (listed above) will have an influence on the resistivity, and hence on resistance. One clear condition that is likely to influence such things as the average distance traveled by an electron between collisions, or the electron density, is a change in temperature. Indeed, it is found empirically that the resistivity of materials does depend on temperature. For metals, the change in resistivity is approximately proportional to the change in temperature, at least near ordinary room temperatures. An adequate approximate expression between resistivity and temperature is the straight line:

$$\rho = \rho_0 \, (1 + \alpha T) \qquad (1\text{-}16)$$

where T is temperature in centigrade, ρ_0, the resistivity at zero degrees and α is called the temperature coefficient of resistivity. Its value for some metals is also given in Table 1-1. Note that the same expression describes the temperature dependence of resistance as can be verified by multiplying both sides by ℓ/A.

Example:

Find the length of aluminum wire having a cross-sectional area of .02 square millimeters which is needed to limit to 100 ma. at 0°C the current drawn from a 12-volt battery (assumed to have zero internal resistance). Also, find the range of the value of resistance during year if the minimum and maximum temperatures are -20° and 35°C. The required resistance is $R = 12/.1 = 120$ ohms. From Eq. (1-15) the required length is $\ell = RA/\rho = \dfrac{120 \times .02 \times 10^{-6}}{2.62 \times 10^{-8}} = 91.6$ meters; the resistivity was taken from Table 1-1. Using Eq. (1-16) for resistance and taking α from Table 1-1 we find at the two extremes of temperature

$$R_{min} = 120 \, (1 - 0.0039 \times 20) = 110.6 \text{ ohms}$$

$$R_{max} = 120 \, (1 + 0.0039 \times 35) = 136.4 \text{ ohms}$$

Table 1-1

	Resistivity in ohm-meters (at 0°C)	Temperature Coefficient of resistivity per degree C at 0°
Silver	1.47×10^{-8}	3.8×10^{-3}
Copper (standard annealed)	1.58	3.8
Aluminum	2.62	3.9
Tungsten	5.5	4.5
Zinc	5.8	3.7
Nickel	6.93	4.3
Iron	8.85	6.2
Platinum	11.0	4.2
Tin	11.5	4.2
Lead	19.8	4.3
Carbon steel	20 to 50	2 to 5
Manganin	43.0	.003
Graphite	800	.075

1.7. Power and Energy

The concept of voltage was introduced by discussing the work done in moving a charge from one point to another. Specifically, the voltage is the work done when a charge moves in an electric field. This work, or energy, is either expended by the charge as it loses potential energy, or it is performed on the charge while moving it to a point of higher potential.

Consider the network branch shown in Figure 1-12. The branch is carrying a current i with a voltage v across its terminals. After the passage of some time, a net charge q will have been transported through the branch from one terminal to the other. From the definition of voltage and the reference directions shown, an energy w is expended by the charge and this energy is

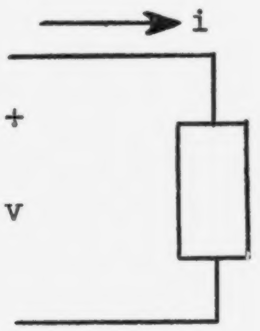


Figure 1-12

$$w = q\,v. \qquad (1\text{-}17)$$

The unit of energy is the joule.

To determine the rate at which this energy is expended, let the charge in question be an incremental charge Δq and let its transfer between the terminals of the branch take place in Δt seconds. Then the incremental work done is $\Delta w = v\,\Delta q$. If we now divide both sides by the time increment Δt and let $\Delta t \longrightarrow 0$, we will find the rate of expenditure of energy, or the power, to be

$$p = \frac{dw}{dt} = v\,\frac{dq}{dt} = vi. \qquad (1\text{-}18)$$

Note again that this is energy expended in the branch by the charge. If either the voltage reference is reversed or the current reference is reversed, work will be done on the charges in moving them to points of higher potential. Thus the power also has a reference direction related to those of current and voltage. The unit of power is the watt, which is the same as a joule per second.

Example

Let the voltage and current in Figure 1-12 be given by the curves shown in Figure 1-13. Find the energy expended at the end of 2 and 4 seconds. Plot also a curve giving the amount of charge passing through the branch as a function of time.

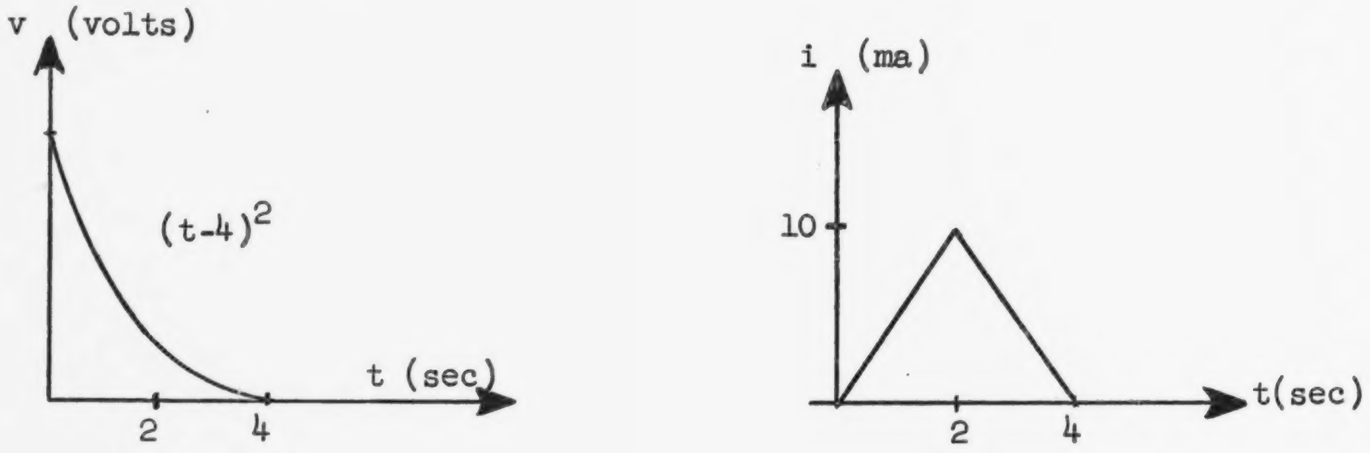


Figure 1-13.

The power is $p = vi$. From $t = 0$ to 2, $i = 5t$ ma. Hence

$$p = 5t\,(t - 4)^2 \text{ milliwatts},$$

and the energy expended after 2 seconds can be obtained by integrating the power.

$$w \text{ (after 2 seconds)} = \int_0^2 5t(t-4)^2 dt = \int_0^2 (10t^3 - 40t^2 + 80t)dt = \frac{280}{3} \text{ millijoules.}$$

For the period from 2 to 4 seconds, it is first necessary to find the equation for i. The straight line has a slope of -5 and passes through the point (4, 0). Hence, $i = -5\,t + 20$. At the end of 4 seconds the energy will be 280/3 plus the integral of the power from 2 to 4 sec.

$$w \text{ (after 4 seconds)} = \frac{280}{3} + \int_2^4 (t-4)^2(-5t+20)\,dt =$$

$$= \frac{280}{3} + \int_2^4 (-5t^3+60t^2-240t+320)\,dt = \frac{280}{3} + 20 = \frac{340}{3} \text{ millijoules.}$$

The charge is the integral of the current. Thus

$$q = \int_0^t 5t\ dt = \frac{5}{2}\,t^2 \text{ millicoulombs} \qquad 0 \le t \le 2$$

$$q = \int_0^2 5t\ dt + \int_2^t t(-t+4)\,dt$$

$$= 10 + 5\left(-\frac{t^2}{2} + 4t\right)\Bigg|_2^t = 10 + \frac{5}{2}\,(t-2)(6-t); \quad 2 \le t \le 4$$

In each interval (from 0 to 2 and 2 to 4 secs.) the curves are parabolas. The complete curve is shown in Figure 1-14. (Verify that the slopes of the two parts are the same at t = 2 and that the slope is zero at t = 4, as they should be.)

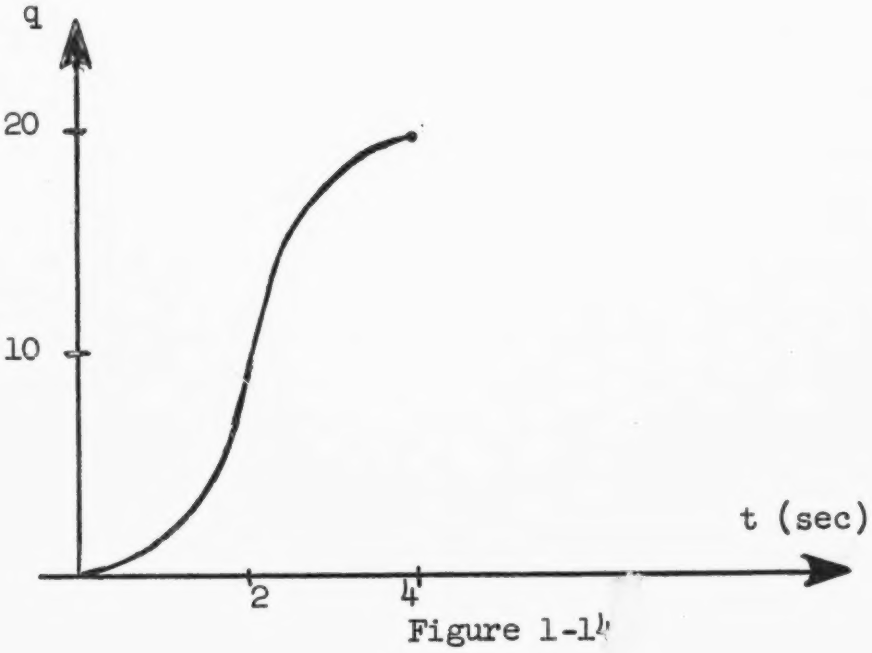


Figure 1-14

If the branch in question is a resistor (ideal) the power expended (this is a short way of saying the rate at which energy is expended) becomes, in substituting Ohm's law into Eq. (1-18),

$$p = Ri^2 = \frac{v^2}{R} = G\,v^2 \qquad (1\text{-}19)$$

Example

Figure 1-15 shows two batteries connected to a 10 ohm resistor. The batteries are each represented by an ideal voltage source in series with an internal resistance. Find the power expended in the 10 ohm resistor and the power supplied by each source. Is the principle of conservation of energy satisfied in this diagram?

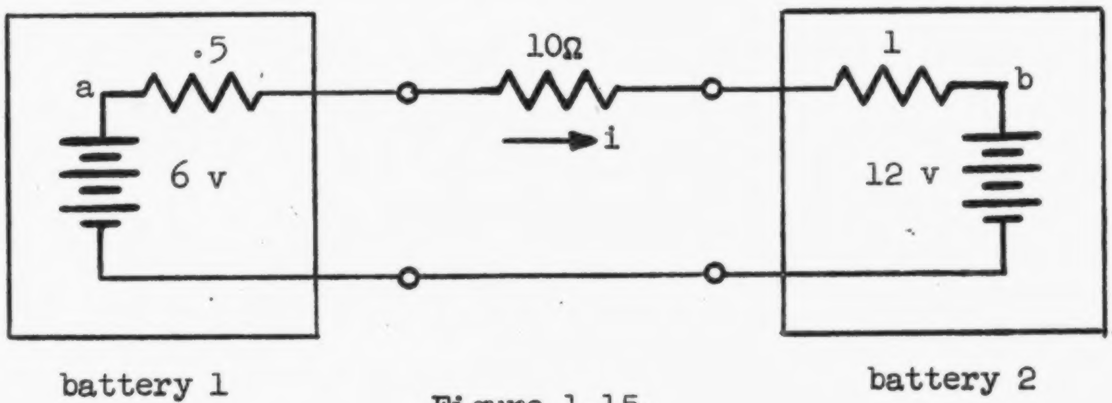


Figure 1-15.

The voltage v_{ab} is 6-12 = -6 volts. This voltage appears across a combination of resistors whose total resistance is 11.5 ohms, so that the current i is -6/11.5 amp. Hence, the power dissipated in the 10 ohm resistor is

$$p = 10\,i^2 = 10\left(\frac{6}{11.5}\right) = 2.72 \text{ watts.}$$

The power entering the ideal 12 volt source is $-12\left(\frac{6}{11.5}\right) = -6.25$ watts. The negative sign indicates that the 12 volt battery is actually supplying power. In the case of the other battery, power entering the ideal 6 volt source is 6(6/11.5) = 3.13 watts. Because the sign is positive, this power is actually absorbed.

To determine the power balance we must also compute the power expended in the two internal resistances as well. To summarize:

power absorbed by 6 volt ideal source = 3.13 watts
power absorbed by 10 ohm resistor = 2.72 watts
power absorbed by .5 ohm internal resistance = 0.14 watts
power absorbed by 1 ohm internal resistance = 0.27 watts
Total 6.26 watts

This is equal to the power supplied by the 12 volt ideal source, as it must be if the principle of conservation of energy is to be satisfied.

Chapter 2

RESISTIVE AND DIODE NETWORKS

In the last chapter three hypothetical devices were introduced, and several "laws" relating to them. There was an ideal resistor, an ideal voltage source and an ideal current source. (The last two will often be abbreviated v-source and i-source.) Kirchhoff's two laws and Ohm's law determine the interrelations of voltage and current in a network containing interconnections of these three devices.

Practical resistance circuits involve the interconnection of devices which, in general, are non-ideal. That is, the v-i curves of resistors are not exactly linear, the potential difference at the terminals of sources is never exactly independent of current (as required for an ideal voltage source) nor is the current of a source exactly independent of voltage (as required for an ideal current source). Nevertheless, there are many cases where resistors have very nearly linear properties, and where actual sources can be represented by equivalent circuits consisting of combinations of resistors and ideal sources. In these cases, actual circuits can be represented on paper by ideal circuits, and their behaviors can be analyzed and predicted by methods developed in this chapter.

We shall discuss procedures developed by the application of these basic relationships in various ways. Our interest will be in computing the voltage or current in a branch of a network, or the power dissipated in a resistor or supplied by a source, when the network itself is given. We are also interested in the converse process, that of determining what a specific resistor value, or source voltage or current, must be in order that a particular branch voltage or current or power take on a specified value. This is a problem in design, or synthesis, as opposed to the previous problem of analysis.

2-1. Series Circuits and Voltage Dividers

A number of branches are said to be in series if they are connected end-to-end such that the current in each branch is the same. Thus, Fig. 2-1 shows a series circuit of three resistors and two voltage sources (one constant, and one variable with time) connected so that the current in each element is the

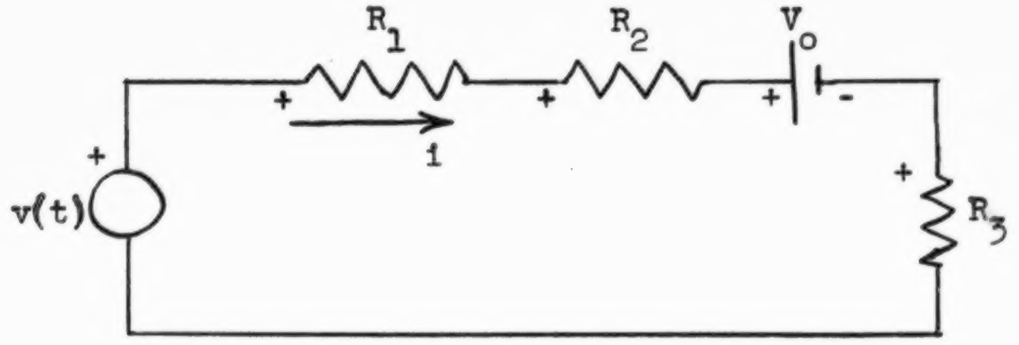


Fig. 2-1 A Series Circuit

same. This relationship of the branch currents automatically satisfies Kirchhoff's current law at the junctions between branches. There is a single closed path around which Kirchhoff's voltage law can be applied. As each voltage term is being written for a resistor, Ohm's law can be applied. With the current reference shown in Fig. 2-1, this simultaneous application of Kvl and Ohm's law leads to

$$R_1 i + R_2 i + V_0 + R_3 i - v = 0 \tag{2-1}$$

which can be solved for the unknown current. Thus,

$$i = \frac{v - V_0}{R_1 + R_2 + R_3} \tag{2-2}$$

Once the current is known, the voltage across any resistive branch follows from Ohm's law.

Now refer to Fig. 2-2, Applying the same technique of analysis, the current

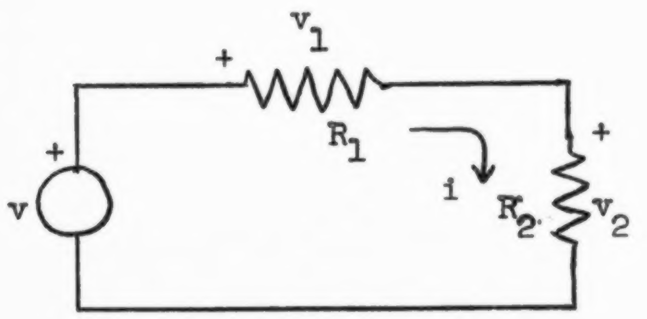


Fig. 2-2 A Voltage Divider

is easily found to be $i = v/(R_1 + R_2)$. The voltage across each of the resistors, with the references shown, can be written

$$v_1 = \frac{R_1}{R_1+R_2}v \quad ; \qquad v_2 = \frac{R_2}{R_1+R_2}v \qquad (2\text{-}3)$$

The structure shown in Fig. 2-2 is called a *voltage divider*; the branch voltage expressions in Eqs. (2-3) are said to be the *voltage divider formula*. It can be remembered as a proportionality as follows: "*The voltage across one resistor of a series combination is to the total voltage what the value of that resistance is to the total resistance*."

2-2. Parallel Networks and Current Dividers

A number of branches are said to be *in parallel* if their branches are connected so that the same voltage appears across each branch. Figure 2-3 shows a parallel network. This relationship of equal branch voltages automatically satisfies Kirchhoff's voltage law around the closed loops formed by the parallel branches. A single independent relationship is obtained by applying Kcl. As each current term is written for a resistive branch, Ohm's law can be applied. With the voltage reference shown in Fig. 2-3 this simultaneous application of

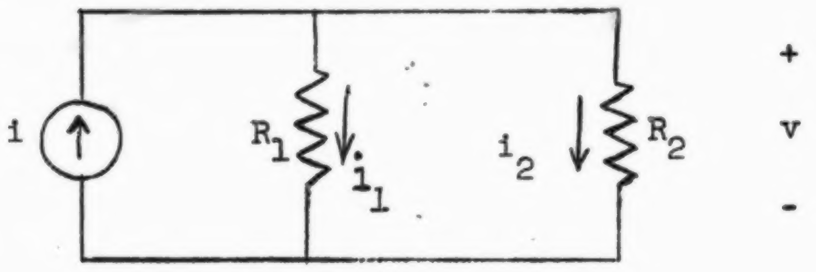


Fig. 2-3 A Current Divider

Kcl and Ohm's law leads to

$$-i + \frac{v}{R_1} + \frac{v}{R_2} = 0 \qquad (2\text{-}4)$$

or, in terms of conductances,

$$-i + G_1 v + G_2 v = 0 \qquad (2\text{-}5)$$

Solving for the voltage leads to

$$v = \frac{i}{G_1+G_2} = \frac{R_1 R_2}{R_1+R_2}\,i \qquad (2\text{-}6)$$

It is seen that the equivalent resistance R of two resistors connected in parallel is given by

$$R = \frac{R_1 R_2}{R_1 + R_2} \qquad (2\text{-}7)$$

In terms of conductances, the equivalent conductance G has the simpler form

$$G = G_1 + G_2 \qquad (2\text{-}8)$$

The current in each resistive branch in Fig. 2-3 is easily found from the voltage in Eq. (2-6) to be

$$\begin{aligned} i_1 &= \frac{G_1}{G_1+G_2}v = \frac{R_2}{R_1+R_2}v && \text{(a)} \\ & && (2\text{-}9) \\ i_2 &= \frac{G_2}{G_1+G_2}v = \frac{R_1}{R_1+R_2}v && \text{(b)} \end{aligned}$$

The structure of Fig. 2-3 is called a current divider. The current divider formula can be easily remembered as a proportionality: "The current in one resistor of a parallel combination is to the total current what the value of that conductance is to the total conductance."

2-3. Network Solution by Equivalent Source Transformations

We have found it a simple matter to find all branch voltages and currents in two network structures: a series circuit and a parallel combination. Suppose a network is given having a structure other than a simple series or parallel arrangement, and that a particular branch voltage or current is to be found. If the structure could be converted to a series or parallel arrangement containing the branch in question, the rest would be simple.

As one step in such a conversion, consider the two networks shown in Fig. 2-4: an ideal voltage source in series with a resistor and an ideal current source in parallel with a resistor. It is assumed that there is a branch (not shown) connected between terminals a and b in both cases, so that there is a current i and a voltage v at these terminals. It is desired to find the conditions on v_0, i_0, R_0 and R for these two configurations to be equivalent at the terminals. By this is meant that

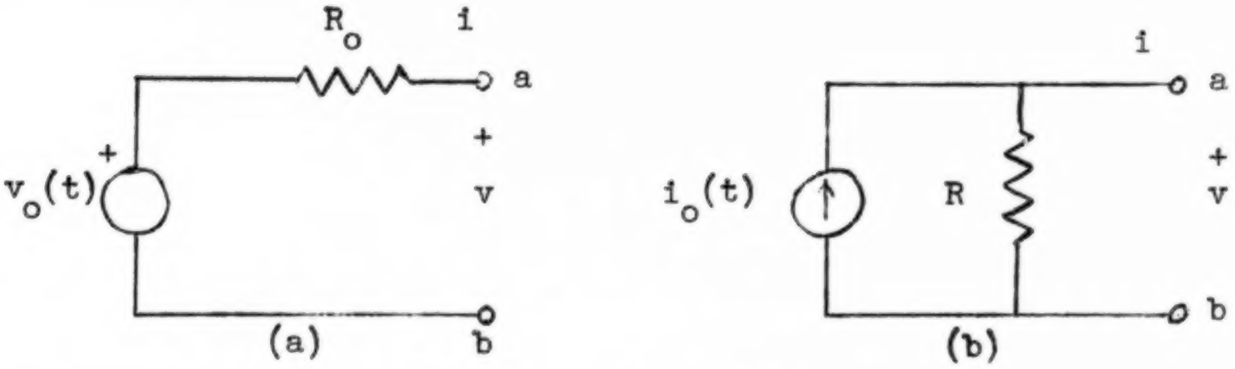


Fig. 2-4 Equivalent Sources

the relationship between the terminal voltage and current is to be the same for the two networks, independently of the load connected at the terminals.

Applying Kvl and Ohm's law in Fig. 2-4a, and Kcl and Ohm's law in Fig. 2-4b, there results

$$v = v_0 - R_0 i \qquad \text{(a)}$$
$$i = i_0 - \frac{v}{R} \quad \text{or} \quad v = Ri_0 - Ri \qquad \text{(b)} \qquad \text{(2-10)}$$

Assuming identical loads, the two voltages should be equal if the two networks are to be equivalent. Equating them leads to

$$(v_0 - Ri_0) + i(R-R_0) = 0 \qquad \text{(2-11)}$$

If the equivalence is to be independent of the load connected at the terminals, this relationship must be valid for all values of i. This will be true only if

$$R = R_0 \qquad \text{(a)}$$
$$v_0 = R_0 i_0 \quad \text{or} \quad i_0 = v_0/R_0 \qquad \text{(b)} \qquad \text{(2-12)}$$

That is to say, the two configurations in Fig. 2-4 are equivalent if the two resistors are the same and the voltage source is related to the current source by $v_0 = R_0 i_0$. Figures 2-4a and b are respectively called a voltage source equivalent, and a current source equivalent. Note that these terms apply to the ideal source together with the resistor, not the ideal source alone. Reference may be made to Chapter 1, to provide a reminder as to how these equivalents relate to actual sources.

With this equivalence, it is possible to reduce a given network to a series circuit or a parallel combination. The process will be illustrated by means of the network in Fig. 2-5. It is desired to find the voltage v across the 20-ohm resistor. The approach will be to convert the network to a series circuit containing the 20-ohm resistor.

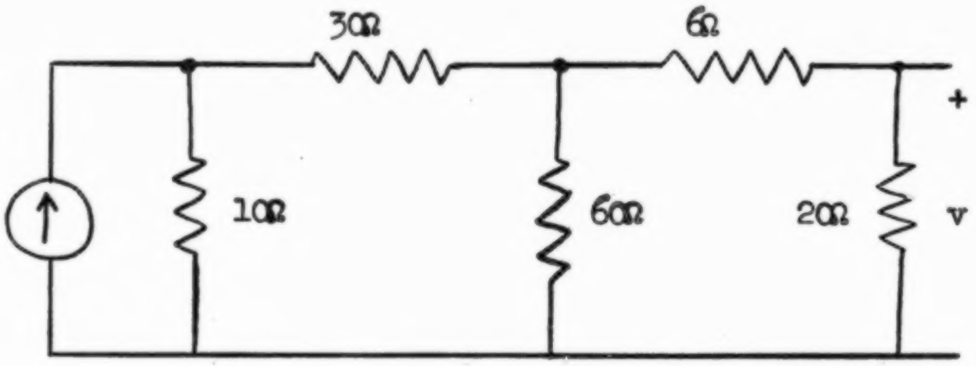
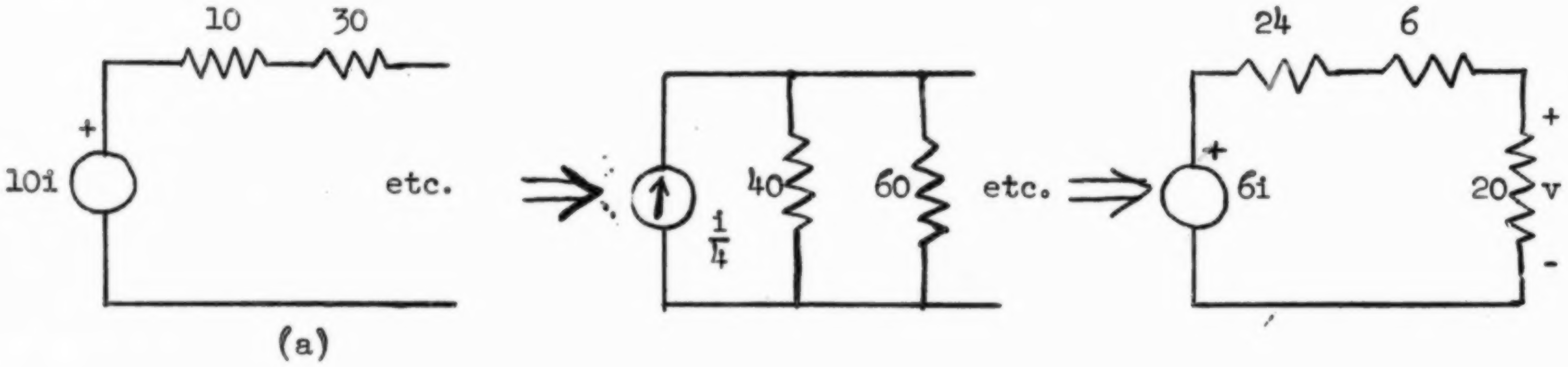


Fig. 2-5

The first step is to replace the combination of the current source and the 10-ohm parallel resistor by its voltage source equivalent--an ideal voltage source 10i in series with 10 ohms. (If it is confusing to have a voltage source which seems to have a current designation, remember that 10i has the dimensions of resistance times current.) The 10 ohms in series with the 30-ohm resistor gives an equivalent resistance of 40 ohms. The series combination of this 40 ohms and the 10i voltage source is then converted to its current source equivalent, as shown in Fig. 2-6b. The 24 ohms equivalent resistance of the 40 ohms and 60 ohms



(a)

Fig. 2-6

in parallel together with the i/4 current source is now converted to their volta source equivalent, as shown in Fig. 2-6c. Finally, application of the voltage divider formula leads to the desired voltage v.

$$v = \frac{20}{20 + 6 + 24} 6i = 2.4i \qquad (2\text{-}13)$$

Since the structure of the original network is destroyed, it is not possibl from the final form in Fig. 2-6 to determine other branch voltages and currents. However, once the desired voltage has been computed, it is possible to return to the original network to find any other desired voltage or current. Thus, suppose it is desired to find the current in the 30-ohm resistor in Fig. 2-5. With v known, the current in the 20-ohm resistor (v/20) is known. But this is the same as the current in the 6-ohm resistor. The voltage across the 60-ohm resist equals the sum of the voltages across the 6- and 20-ohm resistors (6v/20 + v = 13v/10) by Kvl. Hence, the current in the 60-ohm resistor becomes known by Ohm's law (13v/10 divided by 60). Finally, Kcl gives the desired current in the 30-ohm resistor ($i_{30} = i_{60} + v/20 = 13v/600 + v/20 = 43v/600$).

Returning to Fig. 2-26c, note that everything but the 20-ohm resistor has been replaced by a voltage source in series with a resistor. Although this was demonstrated by an example, it is a general result which can be stated as follow

A network consisting of ideal current and voltage sources and linear resist can be replaced at a pair of terminals by an equivalent consisting of a single voltage source and a single series resistance. This circuit is called a *Théveni* *equivalent*, the source being the *Thévenin equivalent voltage* and the resistor being the *Thévenin equivalent resistance*.

Since it has already been demonstrated that a current source in parallel with a resistor can be made equivalent to a v-source in series with the same resistor, it follows that this configuration (an i-source in parallel with a resistor) can be equivalent to any network of sources and resistors at a pair of terminals. This new configuration is called a *Norton equivalent*. The equivalen is illustrated in Fig. 2-7.

Only one process-converting from one source equivalent to another— has been described here for arriving at a Thévenin or Norton equivalent. However, other methods also exist but we shall not consider them here. The method described here depends on having each voltage source in a network

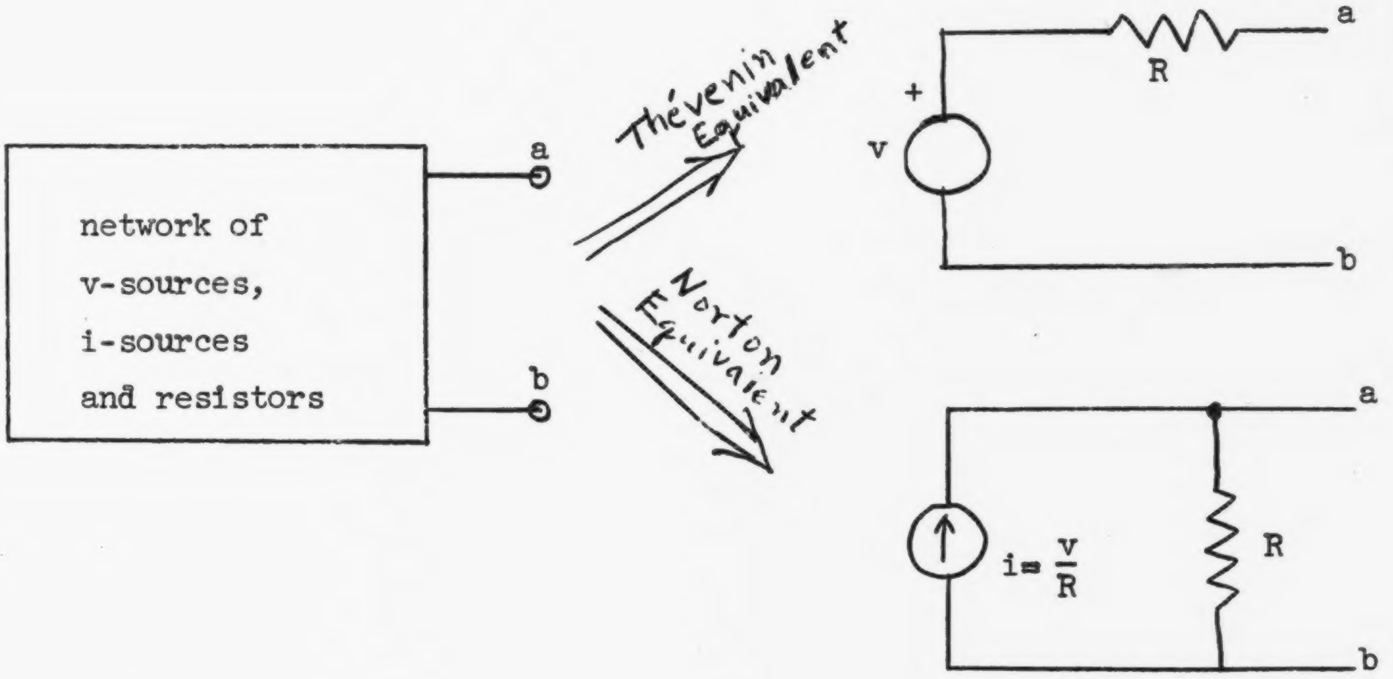


Fig. 2-7

appear with a series resistor and each current source with a parallel resistor. What happens if a source is initially "bare;" that is, no resistor in series with a voltage source or in parallel with a current source? Well, this question does have a favorable answer but discussion of it will be postponed to Sec. 2-6.

2-4. Loop Equations

The preceding method of solving network problems proceeds by converting the structure of the given network into a simple form. We shall now discuss a procedure that uses the three basic relationships--Kvl, Kcl and Ohm's law--in a particular order, thereby arriving at a set of equations. These equations are then solved, thereby determining all voltages and currents in the network. In this method, the structure remains intact,

The procedure will be illustrated in terms of the network in Fig. 2-8(p. 2-10) which is a slight modification of that in Fig. 2-5. The voltage source v_g takes the place of 10i in that network. There are 4 resistors in this network and so 4 resistive branch currents. However, by applying Kcl at the nodes of the network, two of the branch currents can be solved for in terms of the others. (It is trivialy noted that the 6 ohm and 20 ohm resistors are in series so their currents are the same. If Kcl is applied to the node joining these two resistors, the same conclusion will follow.) An expression for the current in the 60 ohm resistor, labeled i_3 in the diagram, is obtained from kcl as $i_3 = i_1 - i_2$.

1 The next step is to apply Kvl around the closed paths, or loops, in
the network. In the present case there are a total of three loops, the
two inner "meshes" and the outside contour, but the Kvl equation for any
one of them can be obtained from the other two, so only two of them are
2 independent. To write Kvl, we need to choose voltage references. Let us
agree to choose all resistive branch voltage and current references with
the plus sign at the tail of the arrow. ( ) Then Ohm's law will
always be written with a plus sign. Now, as we write Kvl around a loop, we
3 mentally replace each voltage by a term of the form Ri with the appropriate
R and i. Thus, writing Kvl around the two inner meshes, we get

4

$$40i_1 + 60(i_1 - i_2) - v_g = 0$$

$$6i_2 + 20i_2 - 60(i_1 - i_2) = 0 \qquad (2\text{-}14)$$

Upon collecting terms and transposing v_g, these become

5

$$100i_1 - 60i_2 = v_g$$

$$-60i_1 + 86i_2 = 0 \qquad (2\text{-}15)$$

6 This pair of linear algebraic equations in two unknowns can be
solved by Cramer's rule in terms of determinants, or by elimination.
The solutions are

7

$$i_1 = .0172\ v_g$$

$$i_2 = .012\ v_g \qquad (2\text{-}16)$$

Once i_1 and i_2 are found, then all other branch currents become known;
8 by Ohm's law, all branch voltages can then be determined. Thus, the voltage
across the 20 ohm resistor will be $20i_2 = .24v_g$. This is to be compared
with the value determined previously in Eq. (2-13), remembering that v_g
here replaces 10i there.

9 The equations that result from this process are called loop equations
since they come from applying Kvl around the loops of the network. The
currents in terms of which the loop equations are written are called the
loop currents.

To summarize: Given a network, first select a number of loop currents and express all branch currents in terms of these loop currents by Kirchhoff's current law. The next step is to write Kvl equations around as many closed paths in the network as there are loop currents while simultaneously substituting Ri's for the v's for each resistance, where each branch current is expressed in terms of the loop currents. The resulting equations are the loop equations. A number of questions present themselves at this point.

1. Which branch currents should be chosen as loop currents and how many?

Except for a single-loop network, the selection of loop currents is not unique and a number of different sets of currents are equally satisfactory. There are well-defined criteria and procedures for selecting an adequate set of loop currents. However, for networks having no more than three loop currents, which is the most we shall encounter, it is actually hard to make a mistake, even if one tries. Hence, no further attention will be given to the subject. It can be proved (although we shall not do so) that the number of independent loop equations in a network having N_b branches and N_n nodes is $N_b - N_n + 1$. This expression can be used as a check to verify that the right number of loop currents have been chosen.

2. Which closed paths should be chosen for writing loop equations?

Here also a number of different possibilities are equally satisfactory. For planar networks (those that can be drawn on a plane without crossing branches) an adequate set of loops are the "meshes", the internal "windows" of the network. Sometimes a different set of loops is more convenient. In any case, there should be as many equations as there are loop currents.

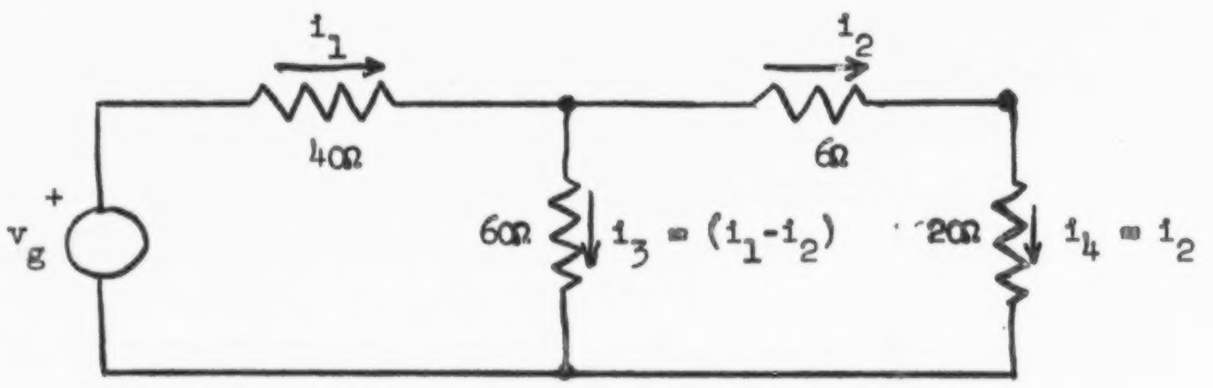


Fig. 2-8

2-5. Node Equations

In writing loop equations, a number of variables--called the loop currents-- are selected, and all branch currents are expressed in terms of these loop currents by Kcl. Let us now, instead, pick a number of voltage variables and express all branch voltages in terms of these variables by Kvl. The process will be illustrated by the network of Fig. 2-5 which is redrawn in Fig. 2-9.

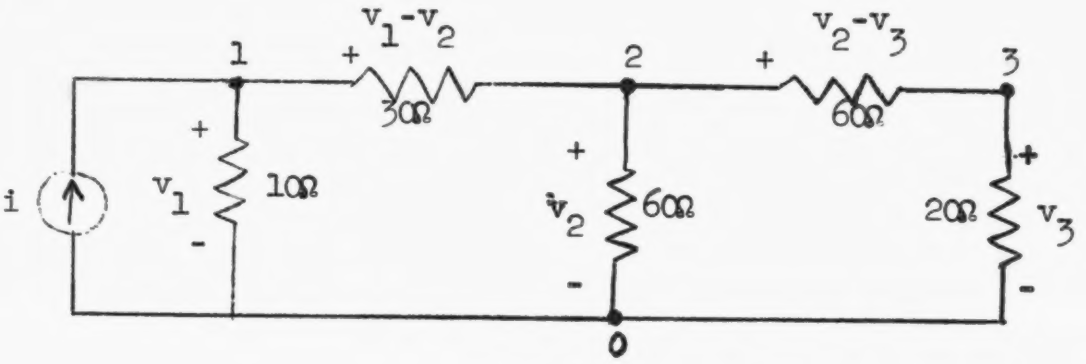


Fig. 2-9

The first step is to choose one node of the network as a datum node to which the voltages of all other nodes will be referred. In Fig. 2-9 let's choose the node labeled 0 as a datum. The voltages of the other nodes relative to that of the datum node, with references chosen plus at the nondatum nodes, are called the node voltages. The voltage of each branch between two nondatum nodes can be written as the difference between two node voltages by Kvl, as shown in Fig. 2-9. Now Kcl is applied at each nondatum node while at the same time replacing the currents by v/R (or Gv), with appropriate v's. The result is a set of equations called the node equations. For Fig. 2-9, the node equations are

$$\begin{aligned}
&\text{node 1:} && -i + \frac{v_1}{10} + \frac{v_1 - v_2}{30} = 0 \\
&\text{node 2:} && -\frac{(v_1 - v_2)}{30} + \frac{v_2}{60} + \frac{v_2 - v_3}{6} = 0 \qquad (2\text{-}17) \\
&\text{node 3:} && -\frac{(v_2 - v_3)}{6} + \frac{v_3}{20} = 0
\end{aligned}$$

Upon collecting terms, clearing fractions and transposing i, these become

$$\begin{aligned} 4v_1 - v_2 \qquad\qquad &= 30i \\ -2v_1 + 13v_2 - 10v_3 &= 0 \qquad\qquad (2\text{-}18) \\ -10v_2 + 13v_3 &= 0 \end{aligned}$$

These node equations are three linear algebraic equations in 3 unknowns and can be solved algebraically. The solutions are

$$\begin{aligned} v_1 &= 8.28i \\ v_2 &= 3.12i \qquad\qquad (2\text{-}19) \\ v_3 &= 2.4i \end{aligned}$$

Note that v_3 is the voltage across the 20 ohm resistor; it was previously labeled v in Fig. 2-5. The answer here agrees with the value found there.

With the node voltages known, all the branch voltages follow; from these the currents can all be determined

To summarize: Given a network, first select a datum node; any node of the network will do. Label the node voltages, which are the voltages of all nondatum nodes relative to that of the datum node. Express all branch voltages in terms of the node voltages by Kvl. Next write Kcl at all the nondatum nodes while simultaneously replacing the currents by voltage-over-resistance, with the branch voltages expressed in terms of node voltages. The resulting equations are the node equations. If there are N_n nodes in the network there will be $N_n - 1$ node equations, all independent.

On comparing the two procedures--loop equations and node equations--we notice that each method utilizes all three of the basic relationships (Kcl, Kvl and Ohm's law) but in a different order. In a network there are initially both current and voltage unknowns. In the case of loop equations, the voltages are all eliminated and expressed in terms of currents; the resulting equations contain only loop currents as unknowns. In the case of node equations, the branch currents are all eliminated and expressed in terms of voltages; the resulting equations contain only node voltages as unknowns.

But it is not essential to follow either of these two methods. One can keep a mixed set of variables--voltage and current--if this should prove more convenient in a given case. We shall not, however, develop detailed procedures for the use of such mixed variables in solving network problems.

*2-6. Additional Comments Concerning Equivalent Sources

The procedure that was used in Sec. 2-3 for obtaining a Thévenin equivalent employs successive conversions from a voltage source in series with a resistor to an equivalent current source in parallel with the resistor, and vice versa. A nagging thought arises here: suppose there is a voltage source without a series resistor or a current source without a parallel resistor, what then? Such a situation is shown in Fig. 2-10(a); no single resistor is in series with the voltage source. Now consider the modification shown in Fig. 2-10(b). The voltage source appears to have slid through the node at its upper terminal into both branches connected there. How are the loop equations modified by this change?

For the loop labeled 3 nothing has been changed so that this loop equation will be the same; only the other two are possibly different. But writing loop equations for the loops labeled 1 and 2 in both the original network and in the modified one in Fig. 2-10(b) (or its redrawn form in (c)) shows that these loop equations are also the same for both cases.

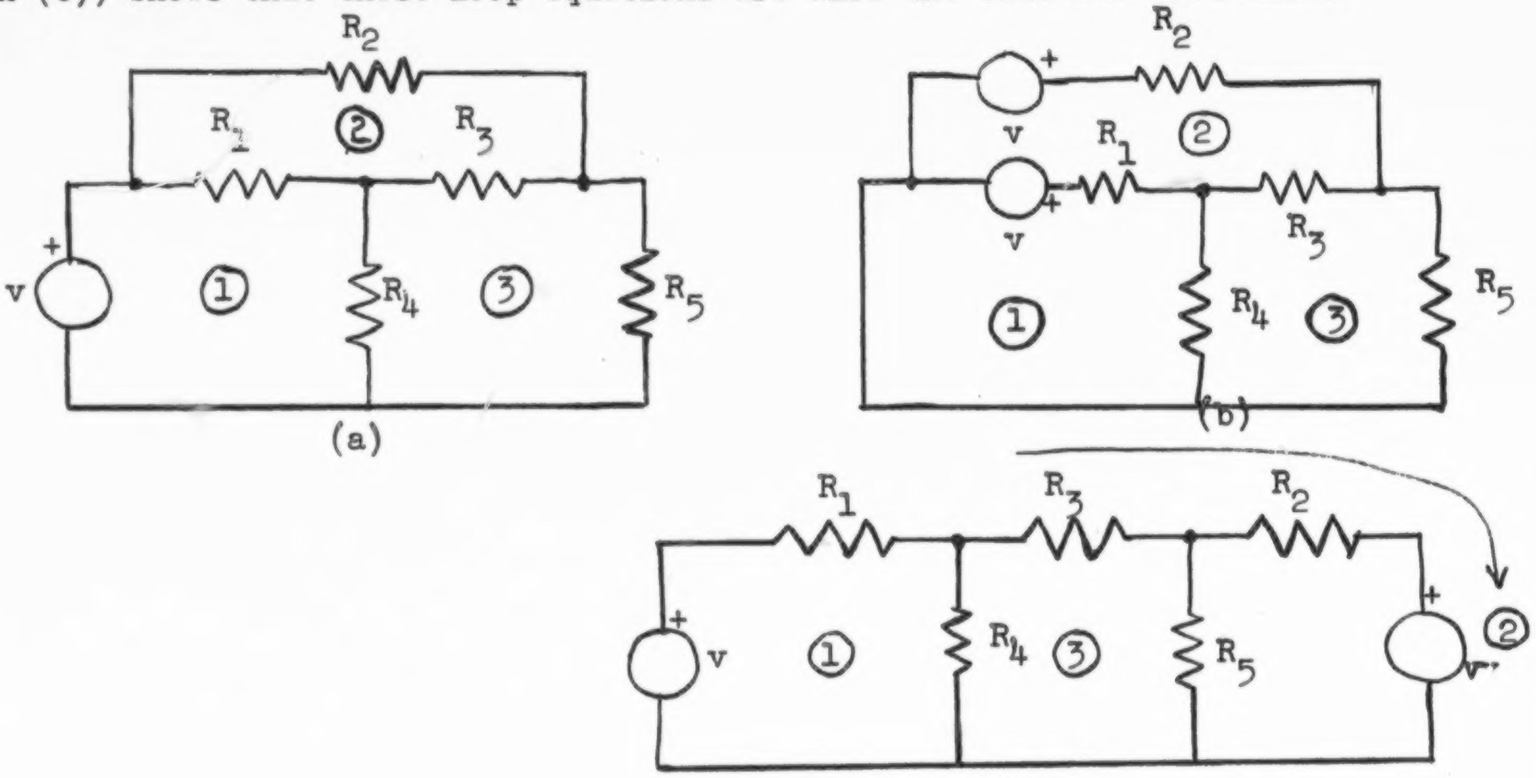


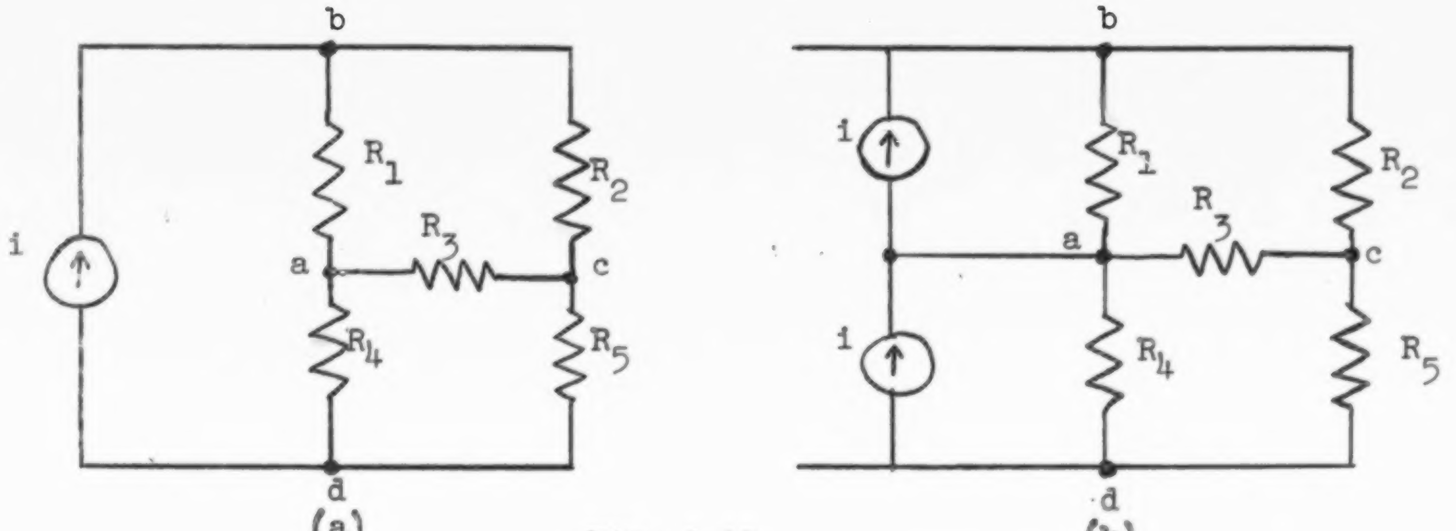
Fig. 2-10

These two cases are then equivalent since they lead to the same values for the loop currents, and thus for all currents and voltages in the network--except for the current in the original source itself. This latter can be easily found by Kcl in original network once all other currents are determined.

This movement of the voltage source has now led to a network having two voltage sources. However, each source is now in series with a resistor and the combination can be replaced by a current source equivalent. Clearly, this result is general; it applies for any number of initially "bare" voltage sources in a network and any number of branches connected at a terminal of each source. It leads to the following general statement

> A voltage source can be moved through one of its terminals into each of the branches connected there, leaving its original position short-circuited without affecting the voltages and currents anywhere else in the network.

How about a "bare" current source, one without an accompanying parallel resistor? Such a case is shown in the network of Fig. 2-11(a) in which the current source i is "bare" and forms a closed path with resistors R_1 and R_4. Now consider the modification shown in Fig. 2-11(b). The current source has been replaced by two sources, both having current i, and their junction has been connected to the node labeled a. There is no current flowing in this connection, as verified by applying Kcl. Hence, the node equation at node a has not been changed. Hence, the two situations in the figure are equivalent, since they will lead to the same node equations and hence, to the



(a)

(b)

Fig. 2-11.

same values of voltage and current--except for the voltage across the original source itself. And this can be easily found from the original network, once all other voltages are known.

Although again the number of sources has increased, each current source has now acquired a parallel resistance, and the combination can be replaced by a voltage source equivalent. Again the result is general and can be stated as follows.

A current source can be moved through any loop it forms with other branches and placed across each of these branches, leaving its original position open-circuited, without affecting the voltages and currents anywhere else in the network.

As a result of these two possibilities concerning the movement and proliferation of sources, even when the sources originally appear "bare" in a network (that is, a v-source without an accompanying series resistor and an i-source without an accompanying parallel resistor) they can be made to acquire accompanying resistive branches, thereby permitting the conversion to an equivalent source.

2-7. The Principle of Superposition

Very often it might be convenient to determine the total current or voltage in a branch of a network containing several sources by finding what this current or voltage would be if each source was the only one in the network, then adding these results. The question is whether such a procedure is valid. The answer to the question is provided by the principle of superposition which is a very general principle applying to a large number of situations in science and engineering. A general statement of the principle is:

Whenever an effect is linearly related to its cause, then the effect owing to a combination of causes is the same as the sum of the effects owing to each cause acting alone, all other causes being inoperative, or deactivated.

In the case of an electric network the effects are currents and voltages in the branches of a network and the causes are the sources. We have seen that any of the equations (loop equations, node equations) that result from applying the basic laws to networks of ideal resistors (and sources) are linear algebraic equations, in which effects are linearly related to causes. Hence, the principle of superposition applies to the calculation of voltage or current in such networks.

1 It only remains to clarify what it means to deactivate a source. A voltage
source is an ideal device which maintains the voltage waveform at its terminals
independent of the terminal current. To deactivate it, or cause it to become
inoperative, means to make its voltage become zero. Zero voltage corresponds
2 to a short circuit. Hence, deactivating a voltage source means short circuiting it.

Similarly, a current source will be rendered inoperative if its current
is reduced to zero. Zero current corresponds to an open circuit. Hence,
3 deactivating a current source means open-circuiting it.

To use the principle of superposition in a network containing several
sources, either voltage or current, all sources but one are deactivated and
a desired branch voltage or current due to the one remaining source is
4 determined. The process is repeated for each source. The sum of the results
due to each source separately are then added to give the total due to all
sources acting together.

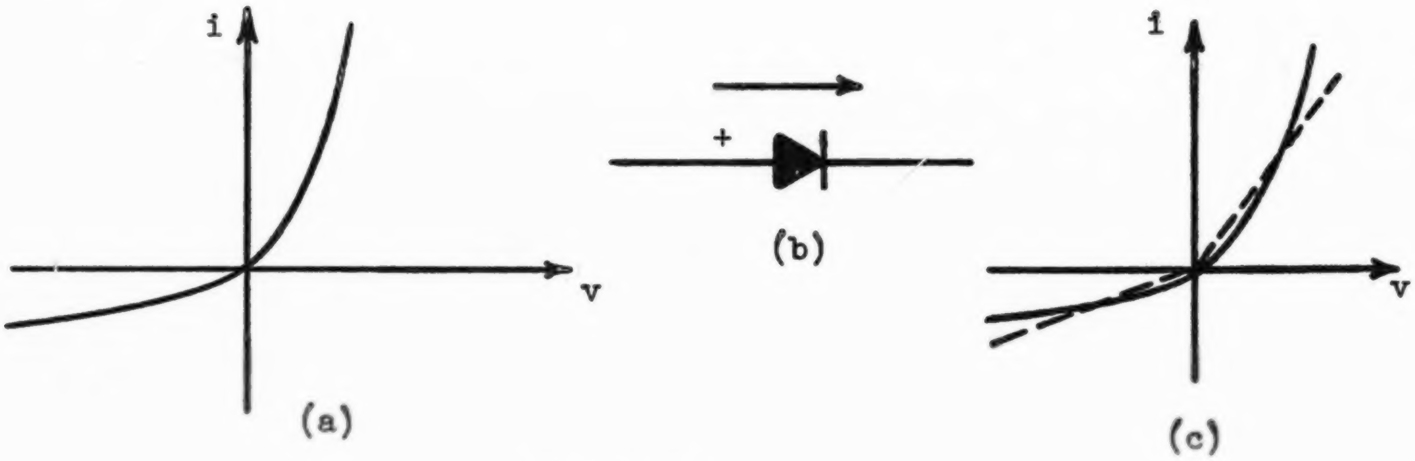
Note well that the principle of superposition is valid only if the effect
5 is linearly related to its cause. Thus, if the desired quantity is the power
dissipated in a resistor due to more than one source, this power cannot be
determined by superposition, since power is not linearly related to current.
Thus, if two sources are present and i_1 and i_2 represent the currents
6 in a resistor R due to each source acting alone, their sum being $i = i_1 + i_2$,
the power dissipated in R when both sources are present is $R(i_1 + i_2)^2$. When
the two sources are acting alone, the sum of the two powers will be
$Ri_1^2 + Ri_2^2 = R(i_1^2 + i_2^2)$. These two expressions are not the same.

7 Similarly, the principle of superposition will not apply in a network
containing devices whose voltage-current relation is not linear, such as
the diodes to be discussed in the next section.

8 2-8. Diode Circuits

A diode is a two-terminal device having a current-voltage curve approximately
like that shown in Fig. 2-12. The symbol used for a diode is also
shown in the figure. A relatively large current in one direction--called
9 the forward direction--is possible with a small voltage. Only a small amount
of current is possible in the other, or reverse, direction.

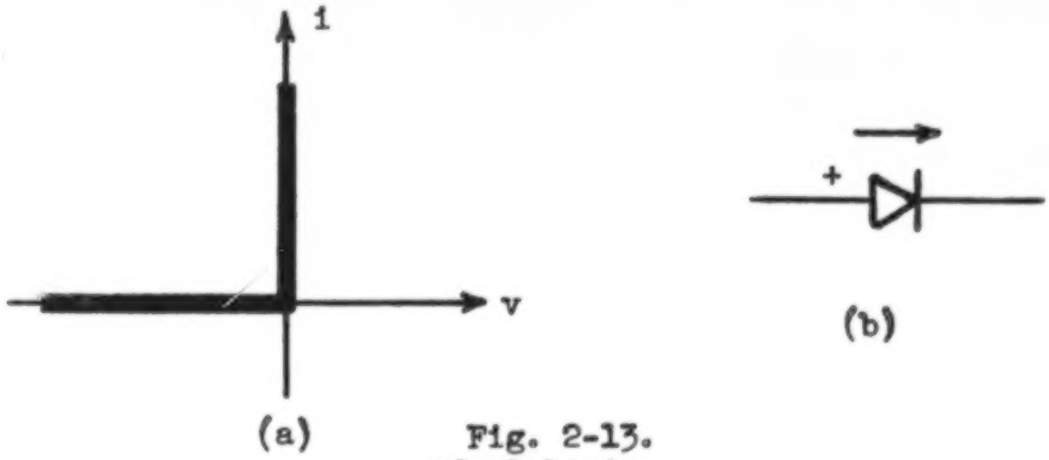
The i-v curve is nonlinear. The greatest curvature, or deviation from a straight line, occurs near the origin, but even at other points the curve is not straight. However, the entire curve can sometimes be approximated by a combination of two straight lines as shown in Fig. 2-12(c). The slope of each line represents a conductance, the reciprocal of a resistance. For positive voltage and current, the resistance is small (large slope) and is called the forward resistance. For negative voltage and current the resistance is large (small slope) and is called the reverse current.



(a) (b) (c)

Fig. 2-12.

It is often convenient to assume that the forward and reverse resistances take on the limiting values: zero forward resistance, $R_f = 0$, and infinite reverse resistance, $R_r = \infty$. The resulting i-v curve is shown in Fig. 2-13. The idealized device having this characteristic is called an ideal diode. To distinguish it from the physical diode, the symbol



(a) (b)

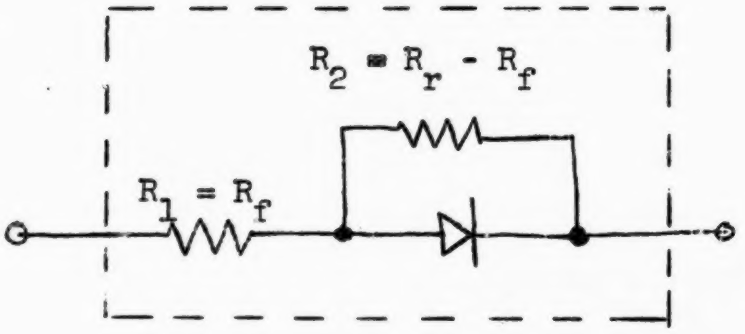
Fig. 2-13.
Ideal Diode.

shown in Fig. 2-13(b) is used. (The arrow head is not black.) The ideal diode has the properties that *forward* (or positive) *current* is accompanied by *zero voltage* and *reverse* (or negative) *voltage* is accompanied by *zero current*. The amount of forward current in an ideal diode is limited by the external network connected at its terminals. The same is true of the amount of reverse voltage.

The ideal diode is seen to be a *two-state* device. When it is conducting, it is said to be *on*; when it is not conducting it is *off*. Whether or not it is in one state or the other is determined by the external network. When making calculations in networks containing ideal diodes, it is not often possible to know beforehand whether a diode is on or off. We *assume* the diode to be in one state or the other, then calculate the diode voltage or current and thereby determine whether the diode is actually in its assumed state. Thus, if the diode is assumed to be *on*, the value of its current can be calculated. If the current turns out to be positive, this verifies that the diode is actually on. If the current turns out to be negative, we conclude that our first assumption about the diode being *on* was not correct; it must have been *off* under the conditions of the problem.

Sometimes, of course, a source voltage or current may be varying with time and so the diode may *switch its state* as the source value changes. Thus, a state of the diode may be assumed, say *off*. With the diode off (open circuited) an expression for its voltage is obtained. From this expression the *critical* value of the varying source voltage for which the diode voltage will turn positive can be determined. A similar condition exists when a network parameter (the value of resistance, say) is not fixed but must be chosen to put the diode in one state or the other.

In many applications sufficiently accurate results are obtained by representing a physical diode as an ideal diode. At other times more accuracy can be obtained if the forward and reverse resistances are not allowed to take on their limiting values. The circuit shown in Fig. 2-14 represents the *piecewise linear model* of a diode, the one having the broken line i-v curve in Fig. 2-12(c). When the ideal diode in this *equivalent circuit* is *on*, R_2 is shorted; hence, R_1 is the forward resistance R_f, a relatively low value. When the diode if off R_1 and R_2 are



Diode Model

Fig. 2-14

in series and together equal the reverse resistance, R_r, a relatively high value.

With an actual diode replaced by a piecewise linear model, the same method of analysis as carried on before is valid. There is the difference, however, that, instead of switching state from open-circuit to short-circuit and back, the overall diode equivalent switches from its reverse resistance to its forward resistance.

When the accuracy provided even by the piecewise linear approximation is not adequate, use must be made of the actual, nonlinear diode characteristic. Consider the circuit shown in Fig. 2-15. The i-v curve of the diode is also shown. The curve provides one relationship between the diode voltage and current. Another relationship is provided by the rest of the circuit. Thus, the voltage across the resistor being V-v, the

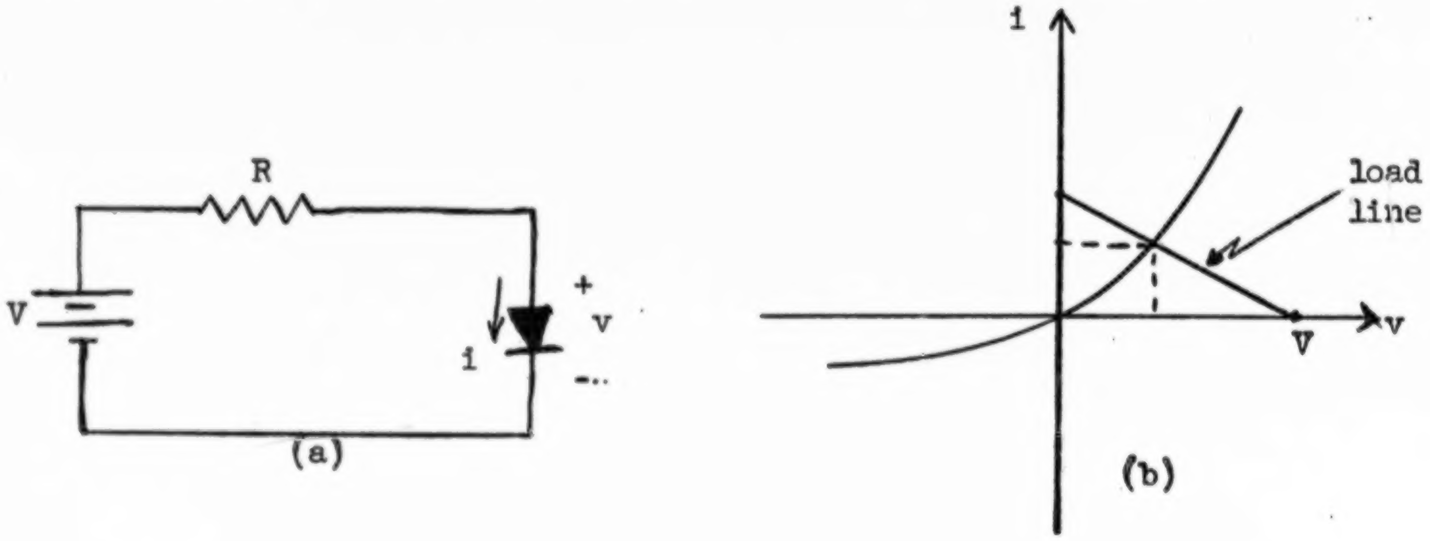


Fig. 2-15

current through it, which is the same as the current i in the diode, is

$$i = \frac{V}{R} - \frac{v}{R} \qquad (2\text{-}20)$$

This is the equation of a straight line whose slope is -1/R and whose intercept on the voltage axis equals the battery voltage. This line is also shown on the same axes as the diode i-v curve in Fig. 2-15(b). It is called the load line. The intersection of the load line with the diode i-v curve gives the solution for the voltage and current of the diode.

The same type of graphical solution can be followed even when the network in which a diode appears is more extensive, containing more ideal resistors and sources, so long as only a single nonlinear diode is present. The rest of the network connected at the terminals of diode can be replaced by a Thévenin equivalent and the result will have the form of the simple series circuit in Fig. 2-15. After the diode voltage and current are determined in the modified circuit, other voltages and currents in the original network can be found by returning to the original network and using the known values of the diode voltage and current.

The graphical load line analysis described here for a network containing a diode can be used in many other cases as well when a single nonlinear device is contained in a "network" of linear devices. This approach will be used later in the analysis ofnonlinear magnetic circuits and of amplifier circuits.

Chapter 3

ELECTROSTATICS

Introduction

The subject of electrostatics has to do with electrical phenomena which can be attributed to the location of stationary charges in space, as distinct from phenomena associated with charges in motion at a constant velocity (magnetism) or accelerating charges (radiation). Electrostatic phenomena never exist completely alone. For example, forces on charged clouds just prior to a lightning stroke are electrostatic forces, and can be understood in terms of the principles of electrostatics. However, the phenomena involved in the assembly of charges on a cloud, and the subsequent lightning stroke are not electrostatic.

Another example of an essentially electrostatic phenomenon is found in the capacitors that are used profusely in electrical circuits. A capacitor charged to a constant voltage comprises a strictly electrostatic situation; it is of limited use or interest. This is the state of a coupling capacitor in an audio amplifier when no signal is being transmitted. In the presence of a signal, however, the voltage and charge are continually changing with time, and a strictly electrostatic situation does not exist. However, at any instant of time, the relationship between voltage and charge is the same as if they were constant (provided their rate of change is not too great) and so electrostatic methods of analysis are appropriate.

Another case, where electrostatic principles apply even though charge is in motion, occurs in the deflection of the electron beam in a cathode ray tube, as the stream of electrons moves between a pair of charged plates. The force on an electron due to these charged plates is independent of the electron velocity, and hence electrostatic principles apply.

3-1. Electric Field

Electrostatic phenomena arise basically from the forces experienced by electrical charges when they are in the presence of other charges. The simplest possible case is illustrated in Fig. 3-1a, showing two small charged spheres separated by a distance r. The charged spheres may be regarded as "small" if their diameters are much less than r.

In terms of this figure, experiment yields the following results:*

(1) The force is one of repulsion if the charges are of like sign;

(2) The force is one of attraction if the charges are of opposite sign;

(3) The magnitude of the force is proportional to the product of the magnitudes of the charges;

(4) The magnitude of the force is inversely proportional to the square of the distance between the charges.

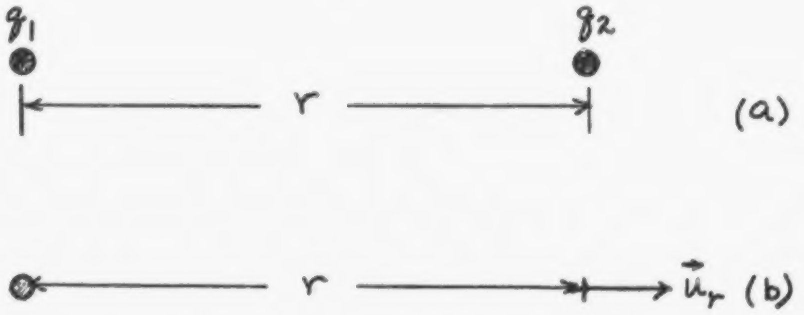


Figure 3-1.

The proportionalities expressed in (3) and (4) are taken care of by using the expression

$$F = \frac{kq_1q_2}{r^2}$$

for the force, where k is a proportionality constant. However, force is a vector quantity and this expression does not account for direction. Let us concentrate on the vector force $\vec{F}$ on q_2. The direction of the vector $\vec{F}$ can be accounted for, in agreement with observations (1) and (2), by using a unit vector $\vec{u}_r$ directed radially away from q_1, as in Fig. 3-1b. Then, the force is completely described by

---

*In these statements it is assumed that means are available for determination of amounts and signs of charges. This is not a trivial question, but is not essential to the description of this basic experiment. Suffice it to say here, therefore, that there are ways to measure amounts and signs of charge.

$$\vec{F} = k\,\frac{q_1\,q_2}{r^2}\,u_r \qquad (3\text{-}1)$$

When q_1 and q_2 are of the same sign, the product q_1q_2 is positive, and the above formula shows $\vec{F}$ in the direction of $\vec{u}_r$. If q_1 and q_2 are of opposite sign, q_1q_2 is negative and so $\vec{F}$ is opposite to $\vec{u}_r$, in agreement with (2).

It has been stated that Eq. (3-1) applies to small charged spheres, without saying how small they should be. It is found that if the diameters are very large (say r/2) then Eq. (3-1) is not valid. In fact, it becomes increasingly accurate as the spheres approach mathematical points. Accordingly, the proper interpretation is that Eq. (3-1) is a postulate, applying to hypothetical point charges. By this we mean that every evidence indicates this relationship is valid, but that it cannot be confirmed by direct experiment, because point charges cannot be attained in the laboratory.

The relationship described by Eq. (3-1) is called Coulomb's Law.

The factor k is experimentally determined, and in the MKSC system of units is found to have the approximate value 9×10^9 (8.988×10^9 is more accurate).* It is a pertinent observation that charge is not defined by this equation, but is defined in terms of current, as its integral with respect to time.

Instead of making a direct substitution of this value of k in Eq. (3-1), for simplification of many subsequent formulas, it is more convenient to replace k by $1/4\pi\epsilon_0$, giving

$$\vec{F} = \frac{q_1\,q_2}{4\pi\epsilon_0 r^2}\,\vec{u}_r \qquad (3\text{-}2)$$

where

$$\epsilon_0 = \frac{1}{36\pi \times 10^9} = 8.85 \times 10^{-12}$$

---

*The experiment described in Fig. 3-1 is not, however, the most accurate way to determine k. One way is to measure the capacitance of an accurately constructed capacitor. Another way, which depends on the development of field theory, is to measure the velocity of propagation of electromagnetic waves (radio or light) in vacuum. Theory shows that $k = c^2 \times 10^{-7}$, where c is the velocity of light.

The quantity ϵ_0 is called the permittivity of free space. The explicit appearance of 4π is entirely arbitrary, being introduced for later convenience.

We now observe that Eq. (3-2) can be written

$$\vec{F} = q_2 \vec{E} \tag{3-3}$$

if the vector quantity E is defined by

$$\vec{E} = \frac{q_1}{4\pi\epsilon_0 r^2} \vec{u}_r \tag{3-4}$$

$\vec{E}$ is called the electric field vector, and is quite fundamental to electrical theory. $\vec{E}$ is the ratio: the force on a test charge q_2, divided by q_2.*

The electric field due to a collection of point charges, like the two charges of Fig. 3-2a, can be determined by applying Eq. (3-4) to each, using respectively r_1 and r_2 for the distance variable r. Vector addition is used, as indicated in the figure. As has been point out, the point charge situation is hypothetical; in all practical cases charge is distributed over a surface (if the charged object is a conductor) or throughout a volume, as in an insulator. Equation (3-4) can be applied to such a situation, to as good an approximation as we like, by imagining the body to be broken up into a system of volume elements, as in Fig. 3-2b, with a point charge at the center of each element. The approximation improves as the number of elements is increased. Thus, in Fig. 3-2b, at a point P the field would be

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\left(\frac{\vec{u}_1}{r_1^2} + \frac{\vec{u}_2}{r_2^2} + \ . \ . \ . \ + \frac{\vec{u}_n}{r_n}\right) \tag{3-5}$$**

where n is the number of elements. Although Eq. (3-5) is given here for

---

*More accurately, the limit of this ratio as q_2 approaches zero. This qualification is necessary because in more general situations, where conductors are present, q_2 will have some effect on the charge distribution on these conductors, and therefore will affect $\vec{E}$ to some extent.

**In field theory, in the limit as each element is reduced to zero this sum becomes an integral.

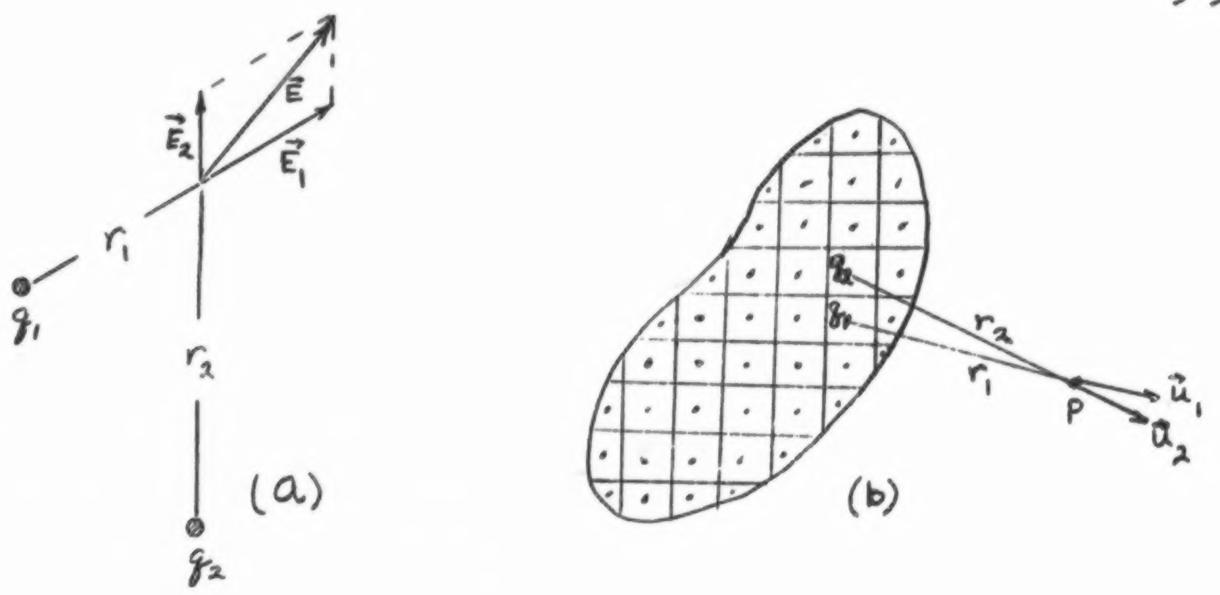


Figure 3-2.

the specific case of a single charged body like Fig. 3-2 , it is a general expression for $\vec{E}$ due to any number of charged bodies, with the understanding that each body is treated in a similar manner. The process of obtaining the $\vec{E}$ vector arising from a distribution of charges by performing a vector addition of the individual contribution of each charge is an example of the principle of superposition. Its validity for finding $\vec{E}$ as described above can be approximately established experimentally with small numbers of charges, but the general applicability to any number of charges, including the distributed charge formulation of Eq. (3-5), is to be regarded as a postulate. As we shall see, this postulate leads to certain theoretical consequences which are amenable to experimental verification.

From the foregoing, it is evident that $\vec{E}$ exists in regions surrounding charged objects, It is sometimes helpful to use sketches of field plots, whereby $\vec{E}$ vectors are drawn at various points in space to show their directions and magnitudes. Some examples are shown in Fig. 3-3. The case shown at (a) is the plot for a point charge; the one in (b) is for a dipole (two charges of equal magnitude and opposite sign spaced a small distance apart). In these plots, vectors are actually shown. In many cases it is sufficient to show only a set of lines indicating the directions of $\vec{E}$ (i.e. radial lines in Fig. 3-3a. The state of the space in

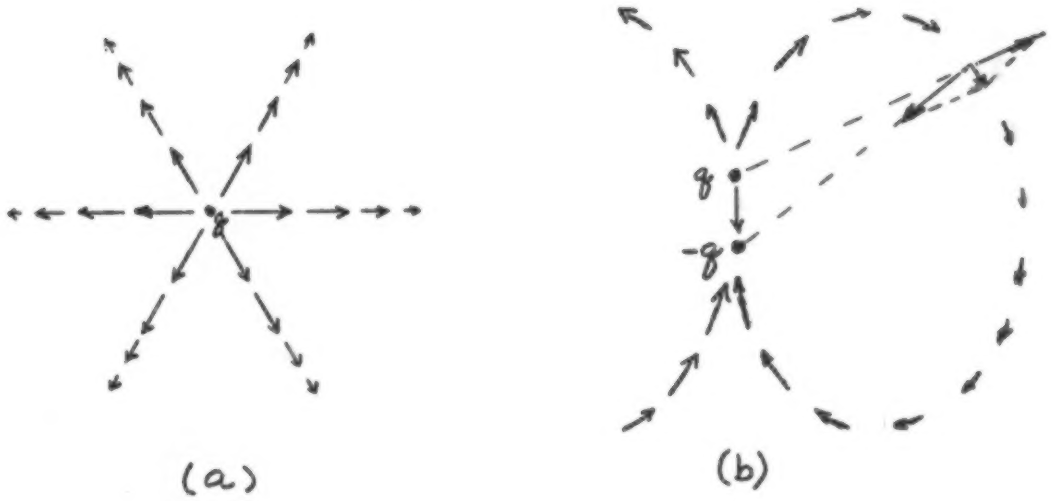


Figure 3-3.

which the field vector $\vec{E}$ is not zero is called an *electric field*.

So far, the discussion of $\vec{E}$ has been in terms of a force vector in free space, where it is easy to imagine measurement of the force on a test charge. A natural question arises as to whether $\vec{E}$ exists within material bodies, and if so, how it can be defined, inasmuch as it is impossible actually to place a test charge in a solid body and measure a force on it. The answer is the rather simple one of using Eq. (3-5) (or the more general integral form referred to in the footnote) to *define* $\vec{E}$ within material bodies. That is, if P in Fig. 3-2b should be inside the body, $\vec{E}$ at that point would be the vector given by Eq. (3-5). There is no inconsistency in doing this; when we come to a consideration of material bodies it will only be necessary to determine what are the consequences of this definition.

Note that an electric field refers to the state where there is a force on a *stationary* test charge. There are situations in which there is a force on a moving charge, due to its motion. This occurs when there is a magnetic field, the subject of the next chapter.

3-2. Properties of an Electric Field

Return to the case of a point charge q, and imagine it to be in the center of an imaginary spherical surface, as in Fig. 3-4. As we have seen,

$$\vec{E} = E\,\vec{u}_r$$

where

$$E = \frac{q_1}{4\pi\epsilon_0 r^2}$$

If we multiply E by the surface area of the sphere, we get

$$4\pi r^2 E = \frac{q_1}{\epsilon_0} \tag{3-6}$$ *

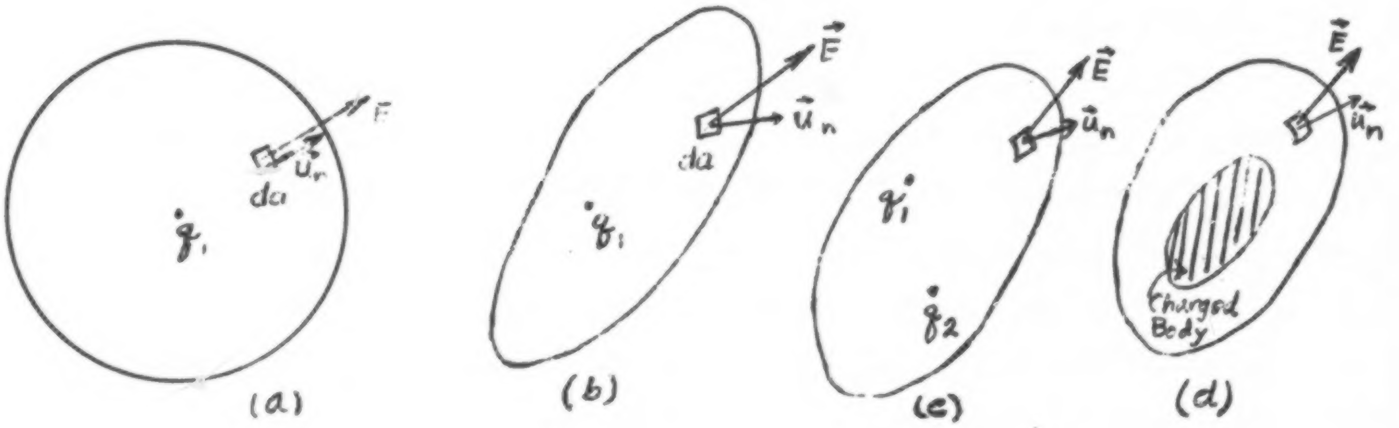

E
$\vec{u}_r$
da
q_1
(a)
$\vec{E}$
$\vec{u}_n$
da
q_1
(b)
$\vec{E}$
$\vec{u}_n$
q_1
q_2
(c)
$\vec{E}$
$\vec{u}_n$
Charged Body
(d)

Figure 3-4.

Now note that the quantity on the left is a special case of the surface integral

$$\iint \vec{E} \cdot \vec{u}_n\, da$$

---

[*] One of the reasons for arbitrarily introducing 4π in Eq. (3-2) was so that the right hand side of Eq. (3-5) would be free of 4π.

over the closed surface of the sphere, where $\vec{u}_n$ is a unit vector normal to the surface (identical with $\vec{u}_r$ in this case). The O through the integral symbols implies a closed surface. Thus, Eq. (3-6) can be written in the more general form

$$\oint\!\!\oint \vec{E} \cdot \vec{u}_n da = \frac{q_1}{\epsilon_0} \qquad (3\text{-}7)$$

which is identical with Eq. (3-6) so long as the surface integral is taken over a sphere of radius r. However, in Eq. (3-7) the surface can be distorted from spherical form, as in Fig. 3-4b. It is not very difficult to show that no matter how the surface is changed, so long as it remains closed with the charge q_1 inside, Eq. (3-7) will always be true.

Now suppose a surface encloses more than one charge, say two, as in Fig. 3-4c. The total vector $\vec{E}$ at a point on the surface is

$$\vec{E} = \vec{E}_1 + \vec{E}_2$$

where $\vec{E}_1$ and $\vec{E}_2$ are due, respectively, to q_1 and q_2. Over the closed surface, then,

$$\oint\!\!\oint \vec{E} \cdot \vec{u}_n da = \oint\!\!\oint \vec{E}_1 \cdot \vec{u}_n da + \oint\!\!\oint \vec{E}_2 \cdot \vec{u}_n da$$

and from Fig. 3-2a it is evident that the two integrals on the right are respectively q_1/ϵ_0 and q_2/ϵ_0. Thus, we get

$$\oint\!\!\oint \vec{E} \cdot \vec{u}_n da = \frac{q_1+q_2}{\epsilon_0}$$

Now suppose a body with distributed charge is enclosed, as in Fig. 3-4d. In Sec. 3-1 it was stated (as a postulate) that bodies with distributed charges can be treated like aggregates of discrete points, to yield Eq. (3-5), and so the above process can be applied repeatedly, each time adding an additional charge represented by a term in Eq. (3-5), to give

$$\oiint \vec{E} \cdot \vec{u}_n da = \frac{q}{\epsilon_0} \qquad (3\text{-}8)$$

where $q = q_1+q_2 + \ldots + q_n$ is the total charge on all bodies inside the enclosing surface. This last equation differs from Eq. (3-7) only to the extent that in Eq. (3-8) the charge is not necessarily at a point.

The statement that Eq. (3-8) is generally valid for all situations of charge distributions, and for all enclosing surfaces, is known as Gauss' law for electric fields. The surface used in an application of Gauss' law is often called a Gaussian surface. Observe that Gauss' law is not a postulate. Although it was not proved here in detail, such a proof is possible, being based on the postulates leading to Eq. (3-5). Parenthetically, it is appropriate to add that one of the purposes in presenting Eq. (3-5) was to make possible the development of Gauss' law. Except in rare instances, Eq. (3-5) is not useful for calculation, but it provides a key step in the development of Gauss' law, which is a very powerful tool in solving problems.

Gauss' law will now be used to investigate a particular situation, that of an isolated spherical shell made of conducting material, of radius R, carrying a charge q, as illustrated in Fig. 3-5. We shall use Gauss' law to determine a formula for $\vec{E}$. Because of symmetry, and the fact that like charges repel, we can conclude that charge is uniformly distributed on the surface. The fact that each particle of charge tries to get as far from its neighbors as possible will prevent any build-up of charge concentration, and will cause the distribution to lie entirely on the surface of the sphere.

We construct a Gaussian surface A of radius r, as shown, and observe that due to symmetry E must be radial at each point on the surface, and therefore can be written $\vec{E} = E\,\vec{u}_r$. Also, observe that $\vec{u}_n$ is also radial (being identical with u_r) so that

$$\vec{E} \cdot \vec{u}_n = \vec{E}$$

Gauss' law gives

$$\oiint \vec{E} \cdot \vec{u}_n da = 4\pi r^2 E = \frac{q}{\epsilon_0}$$

and so

$$E = \frac{q}{4\pi\epsilon_0 r^2} \qquad R < r \tag{3-9}$$

or

$$\vec{E} = \frac{q}{4\pi\epsilon_0 r^2}\vec{u}_r \quad , \qquad R < r \tag{3-10}$$

This equation differs from Eq. (3-4) only to the extend that we have q, the total charge on the sphere, in place of q_1, a point charge, and in the restriction that $r > R$. The reason for this restriction can easily be seen by using another Gaussian surface inside the sphere, labeled B. The charge inside this surface is zero, and therefore at all points on this surface $\vec{E}$ is zero, otherwise the integral of $\vec{E}$ over this surface would not be zero, thereby violating Gauss' law.

Thus we have seen that, for the region of space outside a charged sphere, the electric field is the same as for a point charge located at the center.*

The next step in learning about fields is to consider two concentric spherical shells, as shown in Fig. 3-6a. The inside sphere carries a charge q, as before, and the outside sphere carries a charge q'. For any spherical Gaussian surface lying between the spheres, there is no change from Fig. 3-5. Therefore,

$$E = \frac{q}{4\pi\epsilon_0 r^2} \quad , \qquad R_1 < r < R_2 \tag{3-11}$$

---

*This may seem to contradict an earlier statement to the effect that Coulomb's law does not apply to spheres of finite size. However, here we are considering the electric field due to an isolated sphere, not the force between two spheres. In the presence of another sphere, the charge would not remain uniformly distributed, and this is why Coulomb's law would no longer apply.

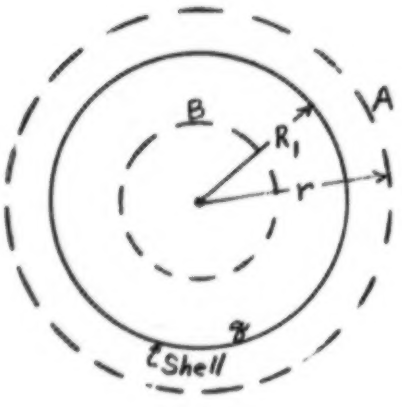


Figure 3-5.

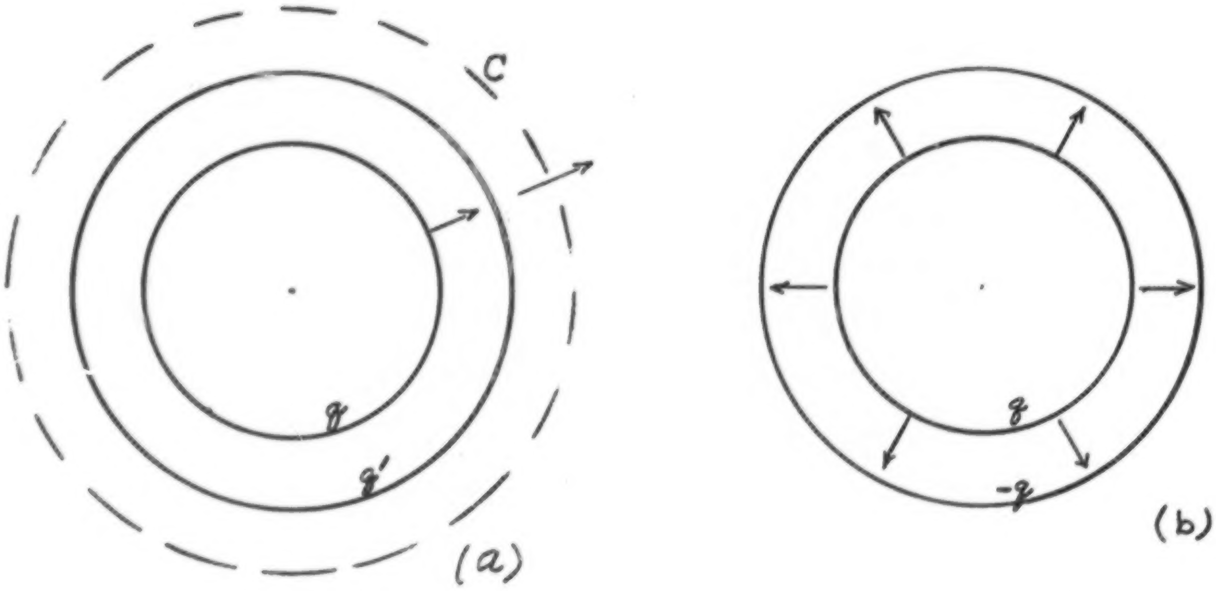


Figure 3-6.

which differs from Eq. (3-10) only in having an upper limit on r. When r is greater than R_2, a spherical Gaussian surface C is used. The charge inside this surface is q + q', and so we have

$$E = \frac{q + q'}{4\pi\epsilon_0 r^2} \quad , \qquad R_2 < r \qquad (3\text{-}12)$$

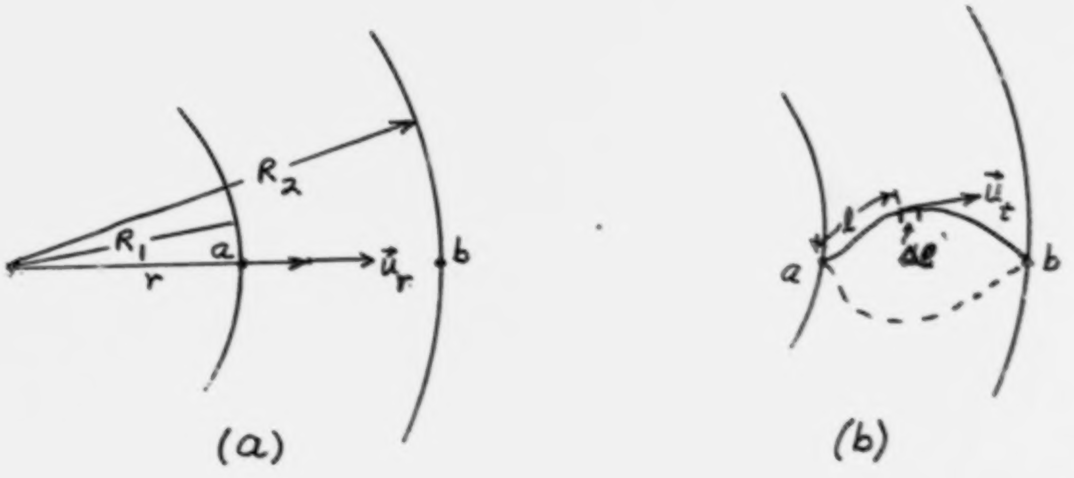
The case where q' = -q is particularly important and interesting. In such a case we learn from Eq. (3-12) that E is zero in the region outside the larger shell. In view of an earlier statement that $\vec{E}$ is zero in the region inside the smaller shell, it follows that for this case the field is confined to the region between the shells, as shown in Fig. 3-6b.

3-3. Potential Difference

Figure 3-7a shows a section of the concentric spheres considered in Fig. 3-6. In this new figure r is an integration variable to be used in calculating the work per unit charge by the field in moving a charge from the inside to the outside sphere. $\vec{E}$ is the force per unit charge, and is everywhere tangent to the straight line path covered by the variable r. Thus, the work is

$$V_{ab} = \int_{R_1}^{R_2} \vec{E}\cdot\vec{u}_r\,dr \qquad (3\text{-}13)$$

Although this is physical work, more specifically it is work per unit charge, and is called potential difference. This is a scalar quantity, requiring the specification of a reference, which is done by calling V_{ab} the potential of point a with respect to point b. This means that the quantity V_{ab} is positive when the field does work in moving a positive charge from a to b, which occurs when the inner sphere is more positively charged than the outer one.



(a) (b)

Figure 3-7.

Equation (3-13) can be evaluated for the case in question, using Eq. (3-10) for E, giving

$$V_{ab} = \frac{q}{4\pi\epsilon_0} \int_{R_1}^{R_2} \frac{dr}{r^2} = \frac{q}{4\pi\epsilon_0} \left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{3-14}$$

We shall have later use for this formula, but the purpose of presenting it here was to show a specific application of Eq. (3-13).

It can be shown mathematically from Eq. (3-5) that the work done by a static electric field in moving a charge between two points is independent of the path taken in going between the points. Thus, rather than the simple straight line path, a curved path might be taken, as in Fig. 3-7b.

For general use, it is convenient to have an integral expression for V_{ab} that can be used for any path. To see what it should be, in Fig. 3-7b let $\vec{u}_t$ be a unit vector tangent to the path, at some point where there is an increment $\Delta\ell$. The vectors $\vec{E}$ and $\vec{u}_t$ are not necessarily in the same direction, and so to get the work done in moving distance $\Delta\ell$ we want the component of $\vec{E}$ along $\vec{u}_t$. Thus, the increment of work is

$$\vec{E}\cdot\vec{u}_t \, \Delta\ell$$

and the potential difference is obtained by summing these in the form of an integral, giving

$$V_{ab} = \int_a^b \vec{E} \cdot \vec{u}_t \, d\ell \tag{3-15}$$

The fact that the above integral is independent of the path between two points is equivalent to a statement that no work is done by a static field when a charge is carried around a closed path. To see this, in Fig. 3-7b, suppose there is a second path, shown by the dotted line. The work done (V_{ab}) by the field is the same in carrying a charge from *a* to *b* by either path. Suppose, however, the charge goes from *a* to *b* along the top route, and is carried back to *a* along the bottom path. The work done in going from *b* back to *a* along the bottom path is the negative of the work done in going from *a* to *b* along this same path.* But this is $-V_{ab}$, and so the work done in covering the closed path is $V_{ab}-V_{ab} = 0$.

A field having the property described above is called a *conservative field*. Since potential difference is what we mean by "voltage" in circuit analysis, it is evident that the above distinctive property of a conservative field is equivalent to Kirchhoff's voltage law.

3-4. Conducting Materials

The idea of potential difference provides a means for defining what we mean by a conductor. A conductor is a material such that under static conditions all points on and within it are at the same potential. In this statement "at the same potential" means that the potential difference between any two points will be zero. Referring to Eq. (3-15), it is seen that this is equivalent to saying that $\vec{E}$ is zero everywhere within a conductor, when charges are motionless.

The fact that $\vec{E}$ must be zero in a conductor has a very interesting consequence. In Fig. 3-8a, the geometrical figure represents a solid conducting body which carries a charge. The dotted figure is a Gaussian surface over which we have

$$\oint\!\!\oint \vec{E} \cdot \vec{u}_n \, da = 0$$

since $\vec{E}$ is everywhere zero. However, from Gauss' law we know that the right hand side of the above equation is the total charge inside the Gaussian surface. The conclusion is that this charge is zero. This analysis applies to *any* Gaussian surface we might choose within the body. Thus, the charge is zero everywhere inside a solid conductor. Whenever a conductor carries a charge, the charge appears on the surface. (Previously we had presented an argument supporting this conclusion only in the case of a sphere.)

---

*It is commonly said that in this case the external force required to move the charge from *b* to *a* does work *on the field*. (i.e. the work done *by* the field is negative.)

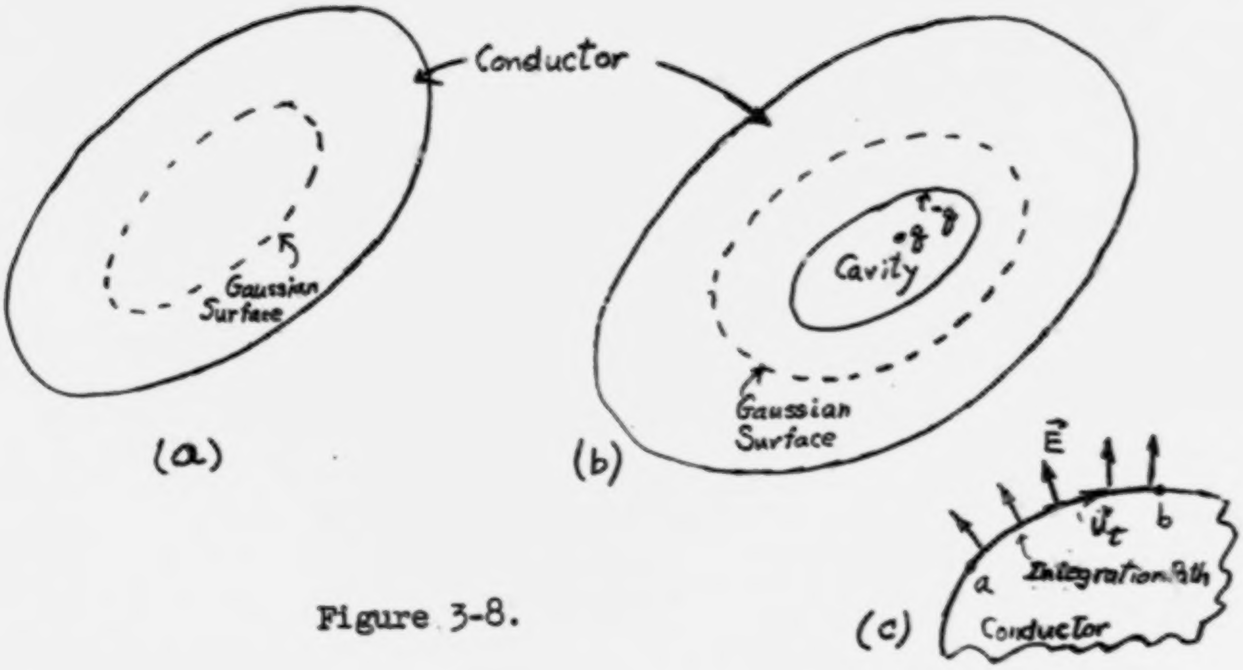


Figure 3-8.

If the conductor is not solid but has an internal cavity as shown in Fig. 3-8b, it is possible that there shall be a charge on the inside surface. This will occur if there is a charged object, insulated from the conductor, inside the cavity. Again using a Gaussian surface, as indicated by the dotted figure, it is known that the charge inside this surface must be zero. Thus, the inside surface must carry a charge -q. If the conducting body is uncharged, there must then be a charge q on the outside surface to make the net charge zero. However, this is not necessary; the outside surface can have zero charge, in which case the conducting body will have a net charge of -q.

The property that $\vec{E}$ is zero in a conductor is a consequence of the existence of "free" electrons, electrons which are free to move about. Thus, any attempt to create an electric field in a conductor results in motion of electrons until they reach a surface and distribute themselves in such a way as to cancel the original cause. This description emphasizes why it is important to say $\vec{E}$ must be zero under static conditions. While electrons are distributing themselves, currents are flowing, and $\vec{E}$ is not zero during the process of charge flow.

Another consequence of the fact that a conducting body is at constant potential is obtained by considering Eq. (3-15) applied between two points on the surface of a conductor, as in Fig. 3-8c. For any two points a and b we must have

$$\int_a^a \vec{E} \cdot \vec{u}_t \, d\ell = 0$$

where the path of integration is along the surface. In general, $\vec{E}$ is not zero, and so the only way for this integral to be zero is for $\vec{E}\cdot\vec{u}_t$ to be zero. Since, $\vec{u}_t$ is tangent to the surface, $\vec{E}\cdot\vec{u}_t$ can be zero only if $\vec{E}$ is normal to the surface, as indicated in Fig. 3-8c.

Since we now know that charge always resides on the surface of a conductor, under static conditions, it can be concluded that in the arrangement of concentric spherical shells described in Fig. 3-7, it makes no difference how thick the shells are. In fact, the inside shell can be replaced by a solid sphere, with no change in the electric field.

We have deduced the property of charge appearing on the surface of a conductor from Gauss' law, which in turn depended upon two postulates concerning $\vec{E}$. At this point it is appropriate to observe that a very accurate experiment is possible to confirm that charge really does reside on the surface.* Thus, this experiment provides a confirmation of the original postulates.

3-5. Capacitance

Return now to a consideration of the two concentric spheres of Fig. 3-7, when the inside sphere carries a charge q, and the outside sphere a charge -q. Equation (3-14) gives

$$V_{ab} = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

as the potential difference between the spheres. Although this equation was derived for the potential difference between points a and b, in the light of Sec. 3-4 it is evident that V_{ab} is the potential difference between any point a on the inside sphere and any point b on the outside sphere.

---

*This is the famuous "ice pail" experiment of Kelvin.

Whenever two charged bodies carry charges of equal magnitude and opposite sign, the potential difference between them is proportional to the charge, with a proportionality factor which depends on the geometrical arrangement. Equation (3-14) is one such example, for the concentric sphere arrangement. This general proportionality can be written

$$V_{ab} = \left(\frac{1}{C}\right)q \tag{3-16}$$

where 1/C is the proportionality factor. C is called the capacitance, and when charge is in coulombs and V is in volts, C is in farads. The dependence of C on geometrical arrangement is illustrated by the present example. From Eq. (3-14) we get

$$C = \frac{4\pi\epsilon_0(R_1R_2)}{R_2-R_1} \tag{3-17}$$

for concentric spheres.

This formula is not of great practical importance because the arrangement of concentric spheres is not practical. However, as we shall see in the next paragraph, it provides a means of analyzing the more practical case of two parallel plates.

We now modify the Eq. (3-14) to give

$$V_{ab} = \frac{q}{4\pi R_1^2\epsilon_0}\left(\frac{R_1}{R_2}\right)(R_2 - R_1) \tag{3-18}$$

and refer to Fig. 3-9 which shows two sections cut out of the concentric spherical shells. This is an hypothetical operation in which it is assumed that the charge distribution is not affected; that is, that the charge that was originally on the cut out sections remains on them. The entire spherical surfaces remain in place, but the two sections are insulated from them. The field distribution will remain undisturbed, and V_{ab}

will be the potential difference between the sections. In the figure, the outside section is only partly shown, but its edges are determined by extending radial lines from the edges of the inside section. Let A be the area of the inside section and q_s its charge, which will be q times the ratio of A to $4\pi R_1^2$, the area of the entire sphere. This calculation gives

$$q_s = \frac{qA}{4\pi R_1^2} \tag{3-19}$$

Likewise, recognizing that the area of the outside section is $(R_2/R_1)^2A$ and that the area of the entire outside sphere is $4\pi R_2^2$, we have

$$\text{Charge on the outside section} = -q\left(\frac{R_2}{R_1}\right)^2 A\left(\frac{1}{4\pi R_2^2}\right) = -\frac{qA}{4\pi R_1^2}$$

The significance of the last calculation is that the charge on the outside section is the negative of Eq. (3-19), and thus that the two spherical

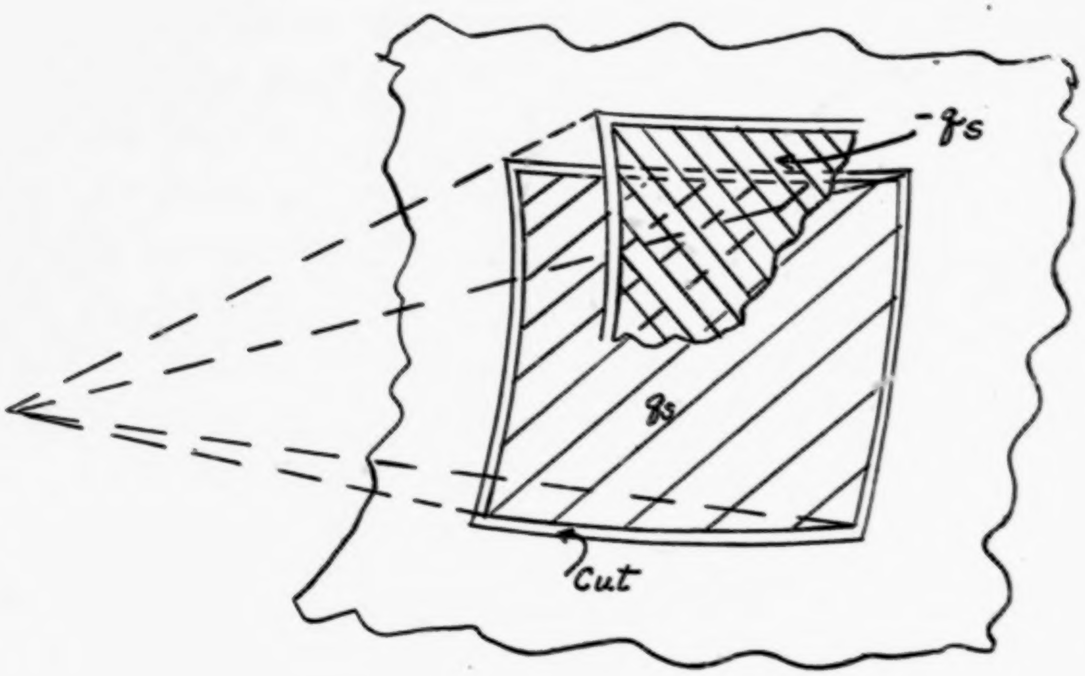


Figure 3-9.

sections have charges of equal magnitude and opposite sign, satisfying the condition for obtaining capacitance between the two sections, as the ratio of charge to potential difference. Equation (3-19) can be solved for q which can be substituted into Eq. (3-18) to give

$$V_{ab} = \frac{q_s}{A\epsilon_0}\left(\frac{R_1}{R_2}\right)h$$

where h has been used for the separation between the shells. Finally, the capacitance is

$$C = \frac{\epsilon_0 A}{h}\left(\frac{R_2}{R_1}\right) \qquad (3\text{-}20)$$

This result can be extended to yield the capacitance between two parallel plane plates by allowing R_1 and R_2 to approach infinity, while keeping $R_2 - R_1 = h$ constant. Geometrically this approaches the situation portrayed in Fig. 3-10a, where the plates are assumed to extend to infinity. In practice, there will be very little change in the central isolated portions if the plates do not extend to infinity. This makes it possible actually to construct the device and to attach wires to it, as in Fig. 3-10b. In this figure, V_{ab} is the battery voltage, and q_s is the integral with respect to time of current i.

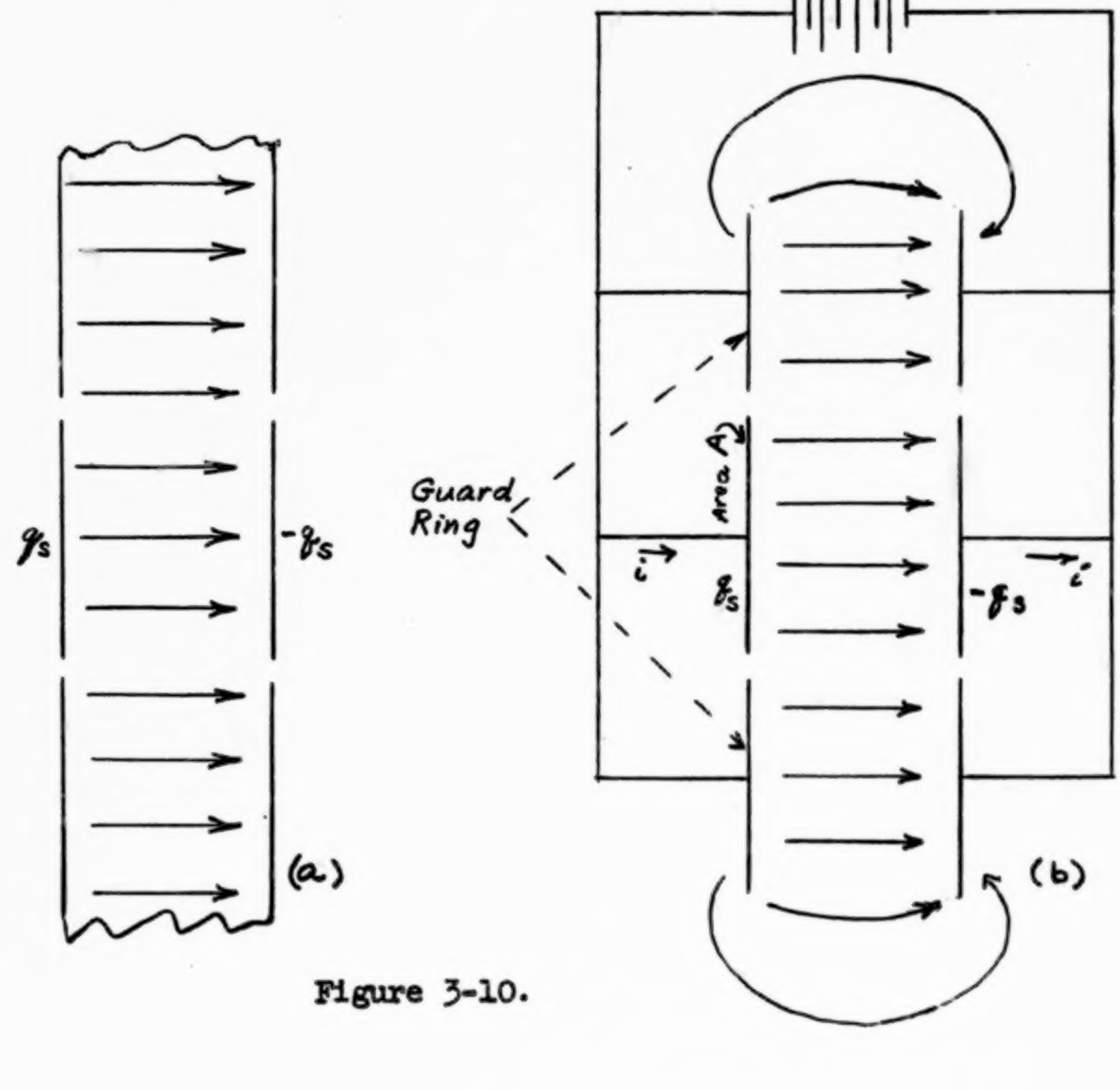


Figure 3-10.

Referring to Eq. (3-20), as R_1 and R_2 both approach infinity, their ratio approaches

$$\frac{R_1}{R_2} = \frac{R_1}{R_1 + h} = \frac{1}{1 + h/R_1} = 1$$

and so Eq. (3-20) becomes

$$C = \frac{\epsilon_0 A}{h} \tag{3-21}$$

for the capacitance between two parallel plates of area A, and spacing h. However, this is accurate only for the arrangement of Fig. 3-10b, where there is a "guard ring" around the plates which serves to keep the field lines parallel between the plates.

We can see that Eq. (3-21) cannot be exactly accurate for an isolated pair of plates, for it requires the E lines to be parallel between the plates, this being the limiting condition of the radial lines for the spherical case. Thus, referring to Fig. 3-11a, assume the plates are isolated and that the lines are parallel between the plates and zero outside. Then the integral of $\vec{E}\cdot\vec{u}_t$ along path (1) would be V_{ab}, but along path (2) the integral would be zero because $\vec{E}$ was assumed to be zero in the region of this path. However, this is impossible, since the potential difference must be the same when calculated by any path. In reality, the field lines must "fringe" out at the edges, as indicated in Fig. 3-11b, in order to make the integral of $\vec{E}\cdot\vec{u}_t$ independent of path. In practice, h is usually very small compared with linear dimensions of the plates, and in that case this fringing has

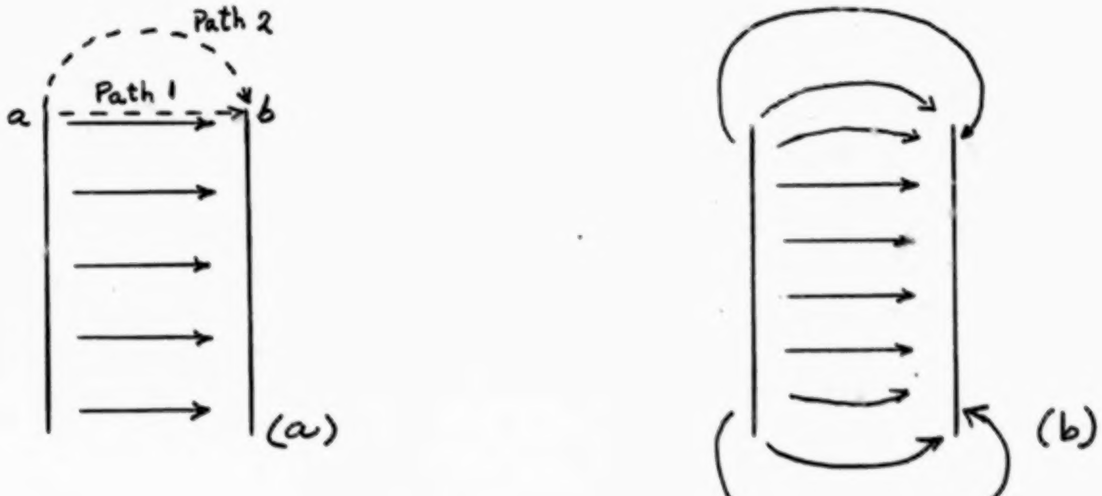


Figure 3-11.

a very small effect, and Eq. (3-21) can be used for the capacitance, to a high degree of accuracy.

3-6. Dielectric Materials

The experiment shown in Fig. 3-10b provides a means for investigating the electrical properties of insulating materials such as oils, glass, paper, plastics, etc. When the space between the plates is filled with one of these materials, it is found that for a given potential difference, the charge (integral of i with respect to t) will be greater than when the space between the plates is occupied by air. Let k_e be the factor by which the charge increases. The capacitance will then be given by

$$C = \frac{k_e \epsilon_0 A}{h} \tag{3-22}$$

The factor k_e is a numerical ratio, and is called the relative permittivity (or dielectric constant). Obviously, the relative permittivity of air is unity.

We shall now consider the consequences of applying Gauss' law to the parallel plate capacitor, when the space between the plates is filled with a dielectric material. Reference is made to Fig. 3-12, where the situations at (a) and (b) are identical, except for the inclusion of dielectric material at (b). Batteries of identical voltage V are used so that $\vec{E} = E\,\vec{u}_n$ (where $E = V/h$) is the same for both. The guard rings force the fringe flux to be far away from the central plates, so that $\vec{E}$ is essentially zero except in the region between the plates.

The dotted rectangle shown in each figure is a side view of a hypothetical rectangular box which serves as a Gaussian surface. Lines labeled ΔA represent surface areas which are normal to $\vec{E}$. Let σ be the density of the charge on the surface of the plate in Fig. 3-12a. Applying Gauss' law to the Gaussian surface in Fig. 3-12a, we observe that the surface integral is zero over surfaces b, c, and d; it is zero over b and c because $\vec{E}$ is parallel to these surfaces, and it is zero over d because $\vec{E}$ is zero on this surface. Thus, Gauss' law gives

$$\oiint \vec{E} \cdot \vec{u}_n dA = (\vec{E} \cdot \vec{u}_n)\Delta A = \frac{\sigma \Delta A}{\epsilon_0} \tag{3-23}$$

and

$$\vec{E} \cdot \vec{u}_n = E = \frac{\sigma}{\epsilon_0} \qquad (3\text{-}24)$$

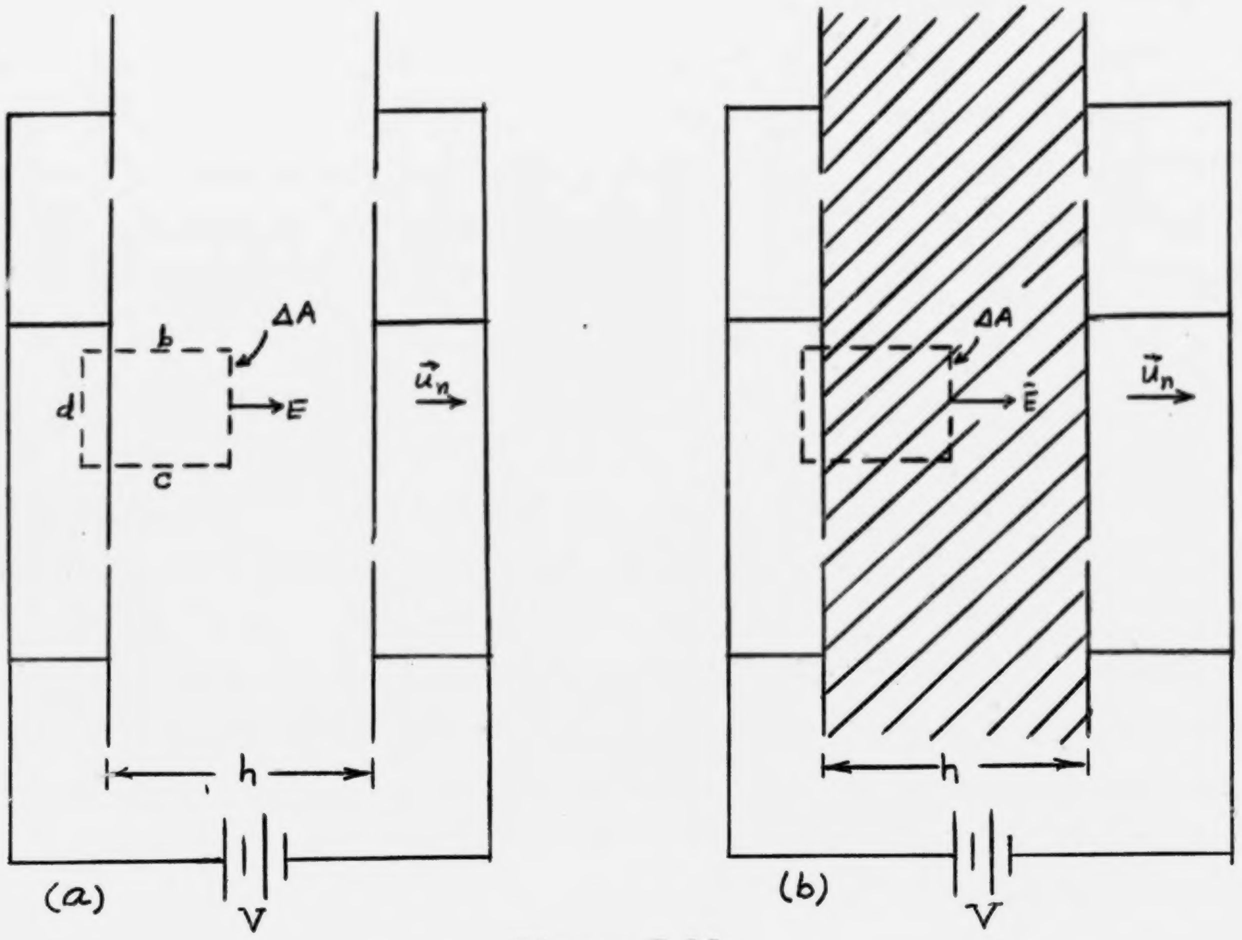


Figure 3-12.

Now observe that E = V/h must be the same for Fig. 3-12b, since V and h are the the same. Thus, knowing E for Fig. 3-12b, Gauss' law can be used to find the charge enclosed. When Eq. (3-24) is substituted into the surface integral of $\vec{E}\cdot\vec{u}_n$, again observing that $\vec{E}\cdot\vec{u}_n$ is zero except on face ΔA, we get

$$\iint \vec{E} \cdot \vec{u}_n da = \frac{\sigma}{\epsilon_0} \Delta A \qquad (3\text{-}25)^*$$

which shows that the enclosed charge is σΔA, and that the surface charge density is σ, the same as in Fig. 3-12a. However, recalling Eq. 3-22, the

---

[*] Equation (3-25) is identical with Eq. (3-23), but is obtained from a different starting point. In Eq. (3-23) Gauss' law is used to obtain E from σ, whereas in Eq. (3-25) Gauss' law is used to find σ from E.

capacitance of Fig. 3-12b must be k_e times the capacitance of Fig. 3-12a, and since $\sigma A = CV$, it follows that the surface charge density plate in Fig. 3-12b should be

$$\sigma' = k_e \sigma \tag{3-26}$$

rather than σ, as predicted by Gauss' law. The conflict implicit in these separate conclusions is resolved by postulating that σ', the charge density on the conducting plate, is modified by a surface charge density σ_b on the surface of the dielectric. Thus, in Fig. 3-12b the total surface to be used in Gauss' law, which we know from Eq. (3-25) to be σ, is also $\sigma_b + \sigma'$. Thus

$$\sigma = \sigma_b + \sigma'$$

and with Eq. (3-26) we have

$$\begin{aligned} \sigma_b &= -(k_e - 1)\sigma && \text{(a)} \\ &= -\left(1 - \frac{1}{k_e}\right)\sigma' && \text{(b)} \end{aligned} \tag{3-27}$$

Since $\sigma = \epsilon_0 E_1$ σ_b, is also given by

$$\sigma_b = -(k_e - 1)\epsilon_0 E \tag{3-28}$$

Experiment confirms this conclusion. Charge σ_b can be explained as follows: due to the force of the E field, the positive and negative atoms of the dielectric are separated slightly, as shown in exaggerated form in Fig. 3-13. This phenomenon results in the appearance of a layer of negative charge on the left hand face, and a similar layer of positive

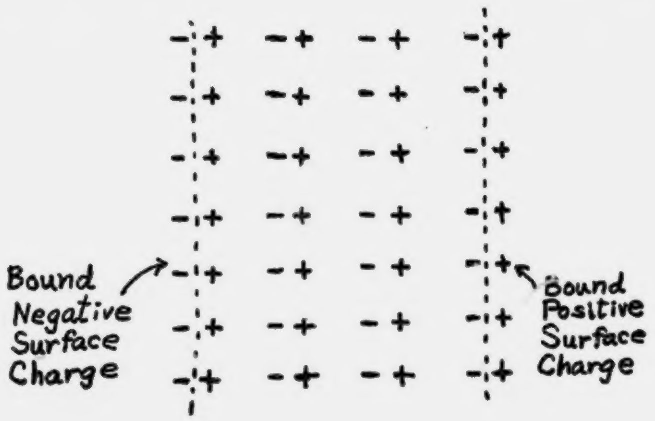


Figure 3-13.

charge on the right hand face. Observe that the negativeness of the charge on the left is in agreement with the negative sign in Eq. (3-25). The atoms represented in this figure are said to be polarized, and the phenomenon is called polarization. Since σ_b is due to displace charges which are parts of neutral atoms of a non-conducting medium, they cannot actually be removed from the surface, say by conduction to a plate. Accordingly, σ_b is called a surface density of bound charge (as opposed to free charge). If the E field is reduced to zero, the positive and negative charges of each atom "spring back" together, and the surface charge disappears.

It appears that Gauss' law, as expressed by Eq. (3-8), is universally valid, even when dielectrics are present, if the charge enclosed within the surface is the sum of all free and bound charges.* It is convenient and useful to have a modified form of Gauss' law in which only free charge will be included. To this end, suppose a new vector

$$\vec{D} = k_e \epsilon_0 \vec{E} \qquad (3\text{-}29)$$

is defined, and consider the surface integral of $\vec{D}\cdot\vec{u}_n$ over the Gaussian surface in Fig. 3-12b. In this case, referring to Eq. (3-24), we have

$$\vec{D} \cdot \vec{u}_n = k_e \epsilon_0 \left(\frac{\sigma}{\epsilon_0}\right) = k_e \sigma = \sigma' \qquad (3\text{-}30)$$

and thus

$$\oiint \vec{D}\cdot\vec{u}_n da = \sigma' \Delta A$$

which is the free charge inside. Although this was shown for a particularly simple case, it can be shown that

$$\iint \vec{D}\cdot\vec{u}_n da = q \qquad (3\text{-}31)$$

is true for any closed surface, where q is the free charge inside. It is immaterial whether or not the surface passes through a dielectric object.

---

*It can be shown that in the general case, where $\vec{E}$ is not necessarily normal to a dielectric surface, Eq. (3-28) becomes

$$\sigma_b = -(k_e - 1)\epsilon_0 E_n$$

where E_n is the normal component of E directed into the dielectric.

Equation (3-30) gives additional information about how D and the free charge are related at a dielectric surface. In terms of Fig. 3-12b, to which that equation applies, we have $\vec{D}\cdot\vec{u}_n = \sigma'$, or

$$D = \sigma' \qquad (3\text{-}32)$$

where $D = \vec{D}\cdot\vec{u}_n$ is the component of $\vec{D}$ inward and normal to the surface. This is generally true, even if the surface is not plane, or if $\vec{D}$ is not normal to the surface. Then the relationship is written

$$D_n = \sigma' \qquad (3\text{-}33)$$

where the subscript (n) is a reminder that this is a normal component. Thus it is seen that $\vec{D}$ has a very simple relationship to the surface density of free charge.

Note from Eq. (3-29) that in air

$$\vec{D} = \epsilon_0\vec{E}$$

so that in air Eq. (3-30) reduces to Eq. (3-8). Thus, Eq. (3-29) is generally valid for all cases, and is a second form of Gauss' law. It is usually the preferred form for the solution of field problems.

The vector $\vec{D}$ is called the dielectric displacement vector. From Eq. (3-30) it is apparent that in the MKSC system of units, $\vec{D}$ is in coulombs/m^2.

3-7. Composite Capacitors

To show that the invention of D is not merely an intellectual exercise, let us use it to determine the capacitance of the parallel plate arrangement shown in Fig. 3-14a. We make the simplifying assumption that the E lines are parallel, as they would be if guard rings were present. Using a Gaussian surface consisting of a rectangular box represented by the dotted rectangle, we note that there is no free charge on the dielectric surface (an assumption of the problem statement). Thus,

$$\vec{D}_1\cdot\vec{u}_n\,\Delta A = \vec{D}_2\cdot\vec{u}_n\,\Delta A$$

and so $\vec{D}$ is the same in air as in the dielectric, and we shall call it $\vec{D}$.

Now, referring to Eq. (3-29), we have

$$\vec{E}_1 = \frac{\vec{D}}{\epsilon_0} \quad , \quad \vec{E}_2 = \frac{\vec{D}}{k_e\epsilon_0}$$

and, finally,

$$V = \frac{h}{2}(\vec{E}_1 + \vec{E}_2)\cdot\vec{u}_n = \frac{h}{2}\left(\frac{1}{\epsilon_0} + \frac{1}{k_e\epsilon_0}\right)\vec{D}\cdot\vec{u}_n \tag{3-34}$$

But $\vec{D}\cdot\vec{u}_n$ equals the surface charge density σ, and the total charge on the surface of area A, is

$$q = \sigma A = (\vec{D}\cdot\vec{u}_n)A$$

Finally, from Eq. (3-34)

$$\frac{h}{2}\left(\frac{1}{\epsilon_0} + \frac{1}{k_e\epsilon_0}\right)\frac{q}{A} = V$$

and the capacitance (q/V) is

$$C = \left(\frac{\epsilon_0 A}{h}\right)\frac{2k_e}{1+k_e} \tag{3-35}$$

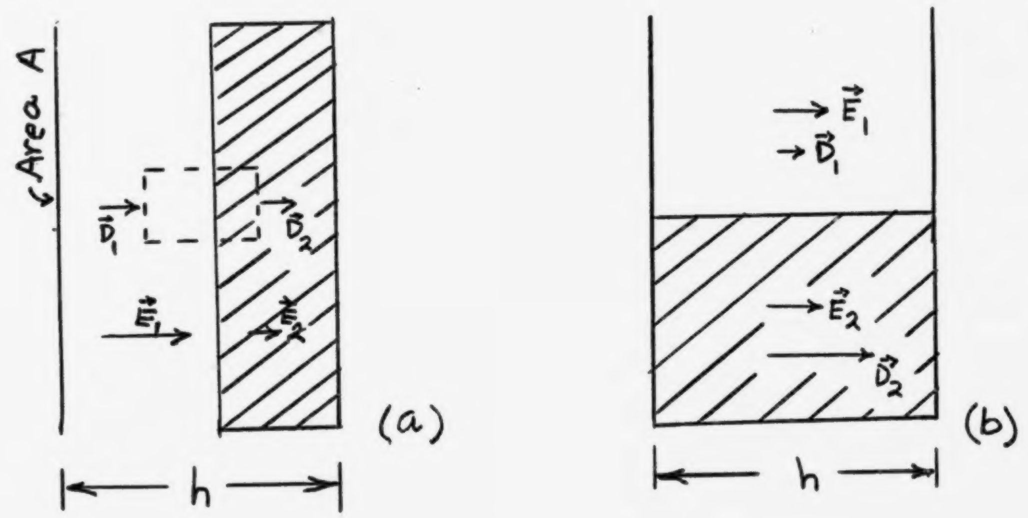

Area A
$\vec{D}_1$
$\vec{D}_2$
$\vec{E}_1$
$\vec{E}_2$
(a)
h
$\vec{E}_1$
$\vec{D}_1$
$\vec{E}_2$
$\vec{D}_2$
(b)
h

Figure 3-14.

Observe that if $k_e = 1$ this reduces to the previous formula for an air capacitor, given by Eq. (3-21), as we should expect.

Another example is shown in Fig. 3-14b. Here the dielectric material extends all the way between the plates, but covers only half of area A. In this case $\vec{E}_1 = \vec{E}_2$, since $(\vec{E}_1 \cdot \vec{u}_n)h = (\vec{E}_2 \cdot \vec{u}_n)h$, by virtue of each plate being at constant potential. Let $\vec{E}$ be this common value of $\vec{E}_1$ and $\vec{E}_2$. Now, in the two sections,

$$\vec{D}_1 = \epsilon_0 \vec{E} \qquad \text{and} \qquad \vec{D}_2 = k_e \epsilon_0 \vec{E}$$

The total charge on the left hand plate, from Eq. (3-21) is

$$q = \frac{A}{2}(\vec{D}_1 \cdot \vec{u}_n + \vec{D}_2 \cdot \vec{u}_n) = \frac{A}{2}(\epsilon_0 + k_e \epsilon_0)\vec{E} \cdot \vec{u}_n$$

But $\vec{E} \cdot \vec{u}_n = V/h$, giving

$$q = \frac{A}{2}\, \epsilon_0 (1 + k_e) \frac{V}{h}$$

and for C we get

$$C = \frac{\epsilon_0 A}{h}\left(\frac{1+k_e}{2}\right) \tag{3-36}$$

It should be fairly obvious that dielectric materials can be useful to provide higher values of capacitance than would otherwise be possible. They do this in two ways; first, by the introduction of the factor k_e in the formula for C, and second, by permitting smaller values of h, for the reason we shall now describe. Dielectric materials, including air, have a property known as dielectric breakdown strength. This is the value of E at which a spark will jump through the material. It is not a precisely determinable quantity because it depends on many factors, such as humidity, smoothness of surface, presence of internal voids, and the like. For air a reasonable figure is 3,000 volts/mm, while solid materials can go as high as 28,000 volts/mm (the handbook figure for polystyrene, a common dielectric material). Thus, if a sheet of dielectric material is used, the plate spacing for a given voltage can be much smaller without fear of dielectric breakdown. Also, reducing h increases C, as can be seen for the cases investigated here, because C is inversely proportional to h. In this way, dielectric materials can be very useful in saving space in capacitor design.

Some typical values of dielectric constants (relative permittivities) are given in the following table.

Material	k_e
Air (760 mm pressure)	1.0006 (usually taken as 1)
Cellulose Nitrate	11.4
Pyrex Glass	4.5
Mica	7.2
Phenol	5.5
Polyethelene	2.26
Polystyrene	2.56
Neoprene	6.7
Benzene	2.15
Petroleum oils	2.2
Ethyl Alcohol	25
Methyl Alcohol	31
Distilled Water	81

3-8. Dielectric Hysteresis

The relationship

$$\vec{D} = k_e \epsilon_0 \vec{E}$$

is approximately true for most materials. However, there are some materials in which there is what might be called a "sluggishness" in the return of the atoms to the unpolarized state when the external polarizing influence has been removed. In other words, the condition portrayed in Fig. 3-13 will persist to some extent, resulting in a partial retention of bound surface charge. Such materials are called electrets. The bound surface charge creates an electric field of its own, and this fact makes it possible to detect this state of "permanent electrification". This phenomenon has possible application for memory devices in computers, although because the similar action in magnetic materials is so much more pronounced, they are more often employed for such a purpose.

The phenomenon just described can be portrayed as a graph of D vs. E, as in Fig. 3-15. If E is increased, starting with unpolarized material, D will increase along the straight line. However, if E is then reduced from point P, D will not return to zero when E is zero, and if E is carried into the negative region, and then back to point P, a closed loop will be formed. This is called a dielectric hysteresis loop. It can be shown that the area of this loop is proportional to the energy lost in the process. This energy loss can be viewed as being due to internal "friction" of the molecules as they react to the changing electric field.

One other possible deviation from Eq. (3-27) should be mentioned. Some substances do not have $\vec{D}$ and $\vec{E}$ in the same direction, and are called nonisotropic.* Advanced mathematics is required to treat this situation, and so no further consideration of it can be given here.

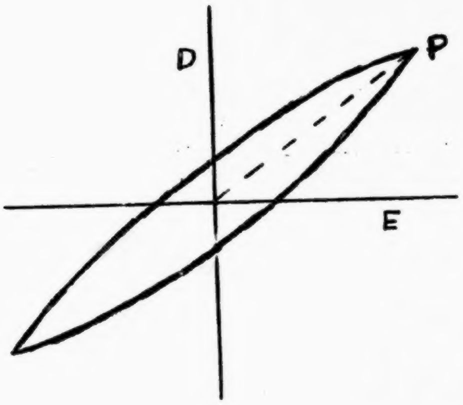


Figure 3-15.

3-9. Resistance Capacitance Circuits

Many practical applications involve circuits which include capacitors. We shall now briefly consider the transient phenomena which arise when an attempt is made to change the voltage across a capacitor. The essential problem is exemplified by the circuit of Fig. 3-16a. A capacitor is uncharged ($v_c = 0$) and then the switch is closed connecting a battery of voltage V, through a resistor R. We are asked to determine how v_c changes with time, subsequent to the closing of the switch. Zero on our time scale is arbitrarily chosen as the instant when the switch is closed.

---

*Nonisotropic materials have different properties along different axes. This also applies to properties other than electrical, such as mechanical deformation and thermal conduction properties.

We recall that the charge q accummulated on the top plate is related to v_c by

$$q = C\,v_c$$

and also that

$$i = \frac{dq}{dt}$$

Thus, i and v_c are related by

$$i = C\,\frac{dv_c}{dt} \qquad (3\text{-}37)$$

After the switch is closed, Kirchhoff's voltage law gives

$$i\,R + v_c = V$$

or, in terms of the variable v_c from the previous equation,

$$RC\,\frac{dv_c}{dt} + v_c = V \qquad (3\text{-}38)$$

This is a differential equation, to be solved for v_c. By writing this as an explicit expression for dt/dv_c we have

$$\frac{dt}{dv_c} = \frac{RC}{V-v_c}$$

From this form we can recognize the antiderivative

$$t = -RC\,\ln(V-v_c) + \ln K \qquad (3\text{-}39)$$

where K is an arbitrary constant. Since v_c is wanted, we can write

$$-\frac{t}{RC} = \ln\left(\frac{V-v_c}{K}\right)$$

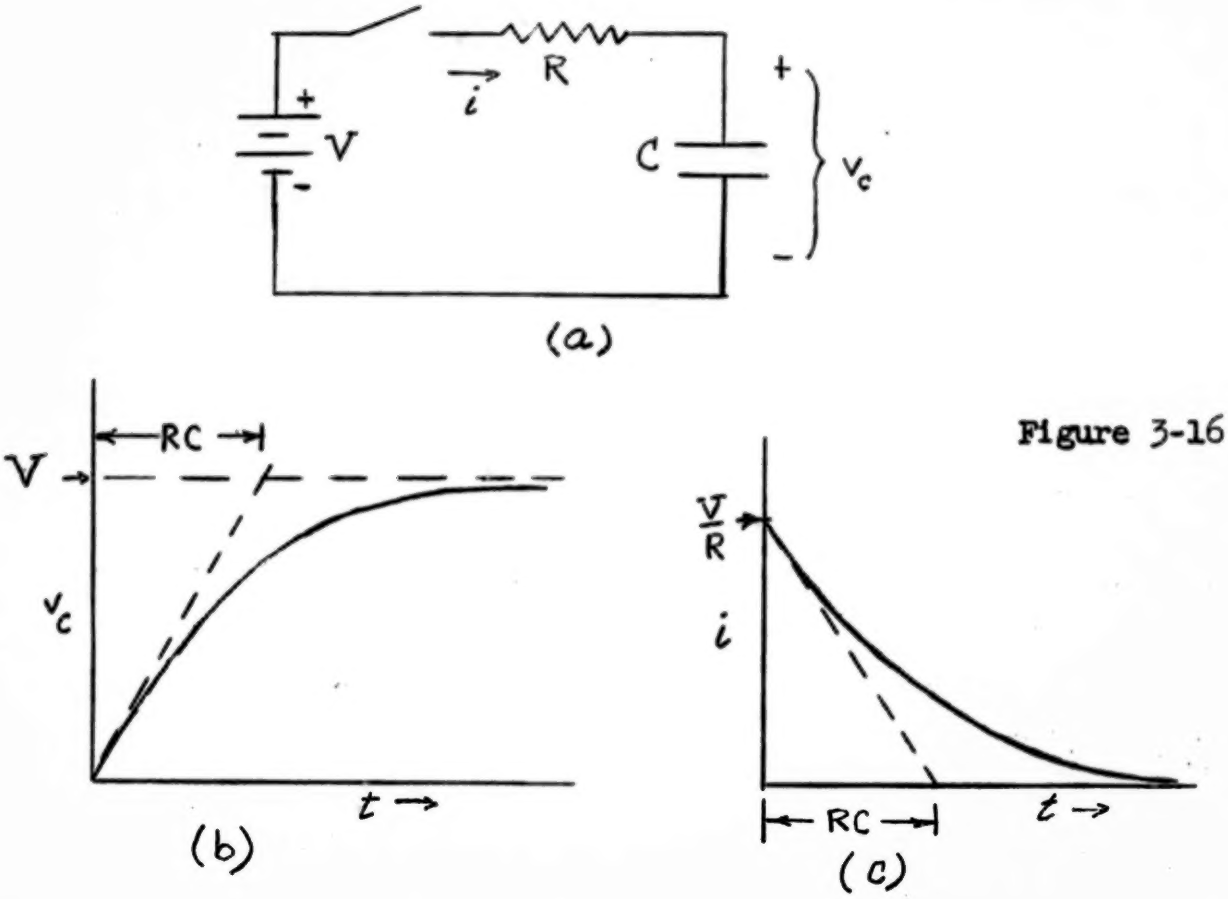


Figure 3-16

and

$$\frac{V-v_c}{K} = e^{-t/RC}$$

or

$$v_c = V - K\,e^{-t/RC} \qquad (3\text{-}40)$$

This is not yet the required solution, because K is an unknown constant which must be evaluated by introducing the initial condition; the value of v_c when $t = 0$. The original statement of the problem gave the information that $v_c = 0$ before the switch is closed. However, Eq. (3-40) begins to apply at the instant when the switch is closed (i.e. for $0 \leqq t$), since it was derived from an equation written for a closed circuit. In connection with this question, observe that if v_c were to experience a sudden jump at $t = 0$, this would mean an infinite derivative (dv_c/dt) and hence Eq. (3-31) shows that the current would be infinite. However, with a finite source voltage, this is impossible, and hence we conclude that a sudden change in v_c is impossible. Thus,

$v_c = 0$ is the appropriate value to use in Eq. (3-40) to correspond to $t = 0$. Substituting these gives

$$0 = V - K$$

and thus the required equation for v_c is

$$v_c = V(1-e^{-t/RC}) \qquad (3\text{-}41)$$

A graph of this function is shown in Fig. 3-16b. This figure includes a geometrical interpretation of the parameter RC, as the time it would take for the voltage to change to its final value, if it continued to change at its initial rate. RC is called the time constant of the circuit, and the final value attained by v_c is called the steady state value.

In addition to learning that in such a circuit the voltage across a capacitor changes exponentially, one of the important observations to be made is that the time required for a change of capacitor voltage to take place depends on the product RC. An interpretation slightly different from the graphical one of Fig. 3-16b involves finding the time interval required for v_c to go through 90% of its total change. This is the value of t at which $e^{-t/RC} = .1$, which occurs when $t = 2.3RC$, approximately. Time constants in practical circuits vary from microseconds, in pulse transmission circuits, to many seconds in certain filter applications.

Equation (3-38) can readily be used to obtain the current,

$$i = C\frac{dv_c}{dt}$$

$$= \frac{V}{R}e^{-t/RC}$$

A graph of this exponential is shown in Fig. 3-16c.

The idea of a change in v_c is further illustrated by the example of Fig. 3-17a. Battery V_1 has been connected for a long period of time so that at $t = 0$ (when the switch is opened) v_c has the value V_1. Opening the switch yields a situation in which the total battery voltage is $V_1 + V_2$, and the total resistance is $R_1 + R_2$. These can be substituted, respectively, for V and R in Eq. (3-40) , to give

$$v_c = V_1 + V_2 - K\,e^{-t/(R_1+R_2)C} \qquad (3\text{-}42)$$

In this case the initial condition is $v_c = V_1$, when $t = 0$, and so

$$V_1 = V_1 + V_2 - K$$

which determines that $K = V_2$. Thus, Eq. (3-42) becomes

$$v_c = V_1 + V_2\left[1 - e^{-t/(R_1+R_2)C}\right] \qquad (3\text{-}43)$$

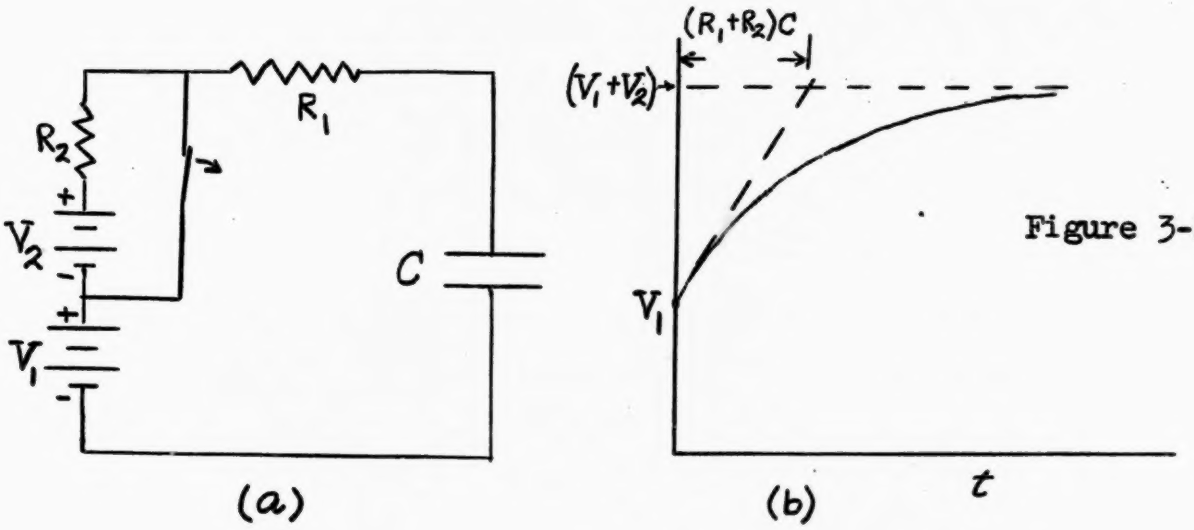
A graph of this function is shown in Fig. 3-17b. The previous interpretation of the time constant still applies, if it is applied to the change of v_c between the initial and final value.

Equations (3-41) and (3-43) are both of the same form, where the change in voltage is represented by a quantity which varies like $(1-e^{-\alpha t})$ X total change of v_c, with the $1/\alpha$ as the time constant. It can be shown that this is typical of the result for any circuit having only one capacitor, and any number of resistors. D-c circuit analysis can be used to find the total change of v_c, and the time constant is C x the equivalent resistance of the circuit connected to C, which results when all sources are reduced to zero.

3-10. Energy Stored in an Electric Field

In Fig. 3-18, a parallel plate capacitor is charging from q_1 to q_2, over a time interval from t_1 to t_2. If q is the charge, and V is the voltage, the current is dq/dt, and thus the instantaneous power input to the capacitor is

$$p = V\,\frac{dq}{dt}$$



(a) (b)

Figure 3-17

The energy supplied to the capacitor is

$$W = \int_{t_1}^{t_2} p\, dt = \int_{t_1}^{t_2} V \frac{dq}{dt}\, dt = \int_{q_1}^{q_2} V dq \qquad (3\text{-}44)$$

In order to convert this to an expression in terms of E and D, we recall that

$$E = \frac{V}{h} \qquad \text{and} \qquad q = AD$$

The second of these expressions is obtained from Eq. (3-32), which is applicable because D is normal to the surface, and q is the free charge (i.e. charge on the plate). Using these in Eq. (3-44), we get

$$W = Ah \int_{D_1}^{D_2} E\, dD \qquad (3\text{-}45)$$

The factor Ah is the dielectric volume, and thus

$$w = \int_{D_1}^{D_2} E\, dD \qquad (3\text{-}46)$$

can be viewed as the density of energy storage in the dielectric.

The integral of Eq. (3-46) is interpreted in Fig. 3-20a, for an electret which is being electrified from an initially unelectrified state (i.e. along the curve which starts from the origin). The change of energy density associated with increasing D from D_1 to D_2 is represented by the shaded area projected back to the vertical axis, as in Fig. 3-19a. Now suppose D is decreased again to D_1. In this case the energy density change decreases by an amount equal to the doubly shaded area. This energy is returned to the circuit, and thus the shaded area between the two curves is energy lost. Similarly, if electrification is carried around a complete loop, as in Fig. 3-19b, the energy lost for this cycle of operation is equal to the loop area. Of course, this energy loss, which is called hysteresis loss, causes heating. In many applications hysteresis loss is negligible in dielectrics

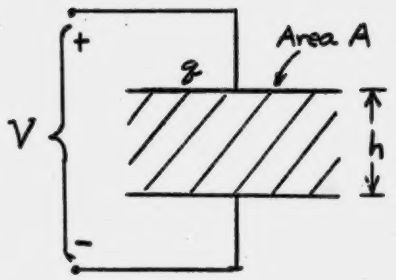


Figure 3-18.

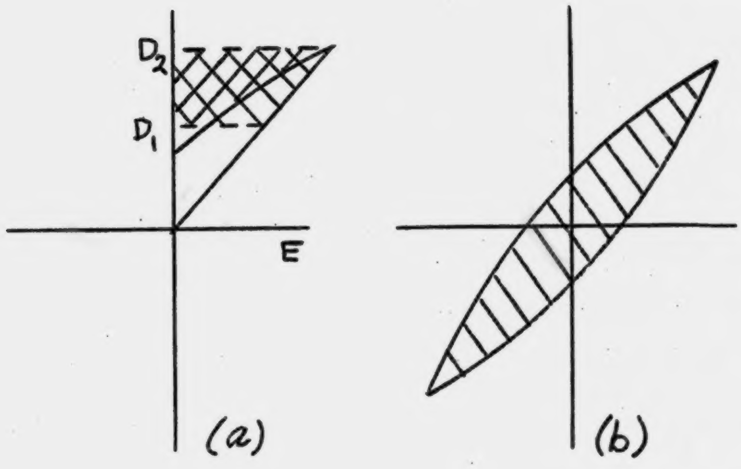


Figure 3-19.

because for most materials the loop area is very small. However, when capacitors are used in a-c circuits, hysteresis power loss can become very high because power loss is energy loss per cycle multiplied by frequency. Thus, a dielectric which behaves with negligible loss at 60 cps. may be totally unsatisfactory at 10^9 cps.

For the important case where $D = k_e \epsilon_0 E$ if we consider the energy increase in bringing D and E from zero to some specific values, from Eq. (3-46) we have

$$w = k_e \epsilon_0 \int_0^E E d\, E = \frac{k_e \epsilon_0 E^2}{2}$$

or

$$w = \frac{DE}{2} \qquad (3\text{-}47)$$

This lost energy relationship is applicable to any dielectric in an electrified state. If it is part of a capacitor the energy can be given in terms of C, as we can see by observing that $i = C\, dv/dt$ so that the integral of the power is

$$W = \int_0^t p\, dt = C \int_0^t v_c \frac{dV}{dt}\, dt = C \int_0^V V\, dV$$

1 or

$$W = \frac{1}{2} C V^2 \qquad (3\text{-}48)$$

2 This energy is completely stored, since in saying that C can be used we have implied that $q = C v_c$ and hence that D is proportional to E, which in turn means there is no hysteresis loop.

3

Chapter 4

ELECTROMAGNETISM

Introduction

In this chapter we shall deal with two basic physical phenomena. First, the phenomenon of mechanical forces (and/or torques) exerted on current-carrying circuits either in the vicinity of another current-carrying circuit, or a permanent magnet. Second, the creation of an electric potential difference as the result of motion of a conductor, or as a result of a changing current in another circuit. The first of these phenomena is the basis of the subject of magnetism. Since the second involves an interrelationship between electrical and magnetic effects, it is called an electromagnetic phenomenon.

Almost every practical electrical device involves electromagnetic phenomena. For example, the electric power we use is generated by the motion of a conductor in a magnetic field, the diaphragm of a telephone receiver is actuated by magnetic forces, magnetically operated relays are used to open and close circuits in many applications such as motor controls and telephone switching circuits, and, of course, electric motors operate on the principle that a force is exerted on a current-carrying conductor in a magnetic field. Thus, these few illustrations indicate that a study of electromagnetic phenomena is important.

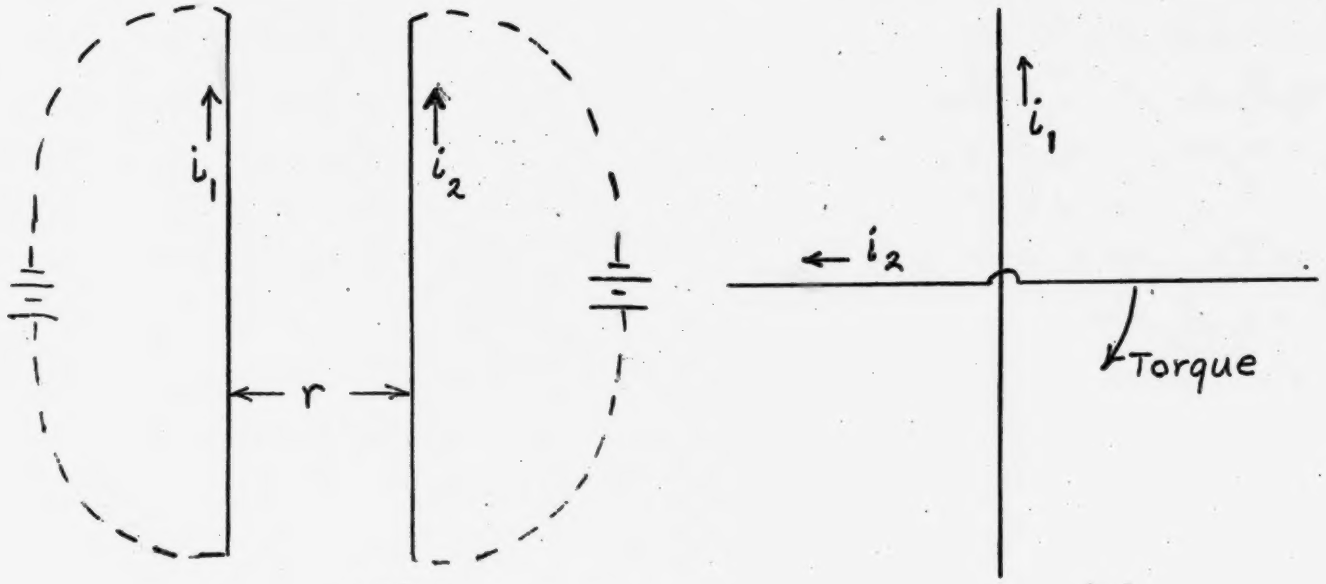
4-1. Basic Magnetic Experiments

Consider an experiment in which two long straight parallel wires carry currents, as shown in Figure 4-1(a). Wire (1) is rigidly supported, and provision is made to measure the force on wire (2). The following observations can be made:

1) When i_1 and i_2 are in the same direction (either both in the reference directions shown, or both opposite), wire (2) experiences a force toward wire (1).
2) If either current is reversed (but not both), the force on (2) will be away from (1).
3) In either case, the magnitude of the force is proportional to the magnitude of the product $i_1 i_2$.

4) The force is proportional to $1/r$, where r is the separation between wires.

A second experiment can be conducted as shown in Figure 4-1(b). Here, the wires are at right angles and in the same plane, one passing in back of the other by virtue of a small semicircular jog in one of them. In this case, there is found to be no net force on wire (2), but there is found to be a torque tending to rotate wire (2) into a position parallel with wire (1) so that i_1 and i_2 will be in the same direction. Thus, for currents in the reference directions indicated in Figure 4-1(b), the torque on wire (2) is found to be clockwise.



(a) (b)

Figure 4-1.

4-2. Flux Density

In Figure 4-1(a), wire (2) experiences a force, and so it is reasonable to assume that there is a force on a short length $\Delta\ell$, as indicated in Figure 4-2(a). With fixed i_1, the force is toward or away from wire (1), depending on the direction of i_2. This much of the experiment suggests, perhaps, that the force on $\Delta\ell$ can be described by defining a vector directed radially from wire (1). However, when we go to the crossed wires, as in Figure 4-2(b), we see that in order to provide the observed torque, the force on an element $\Delta\ell$ must be parallel to wire (1); up or down depending on the direction of i_2. The force on the element $\Delta\ell$ is always perpendicular to the direction of current flow in $\Delta\ell$.

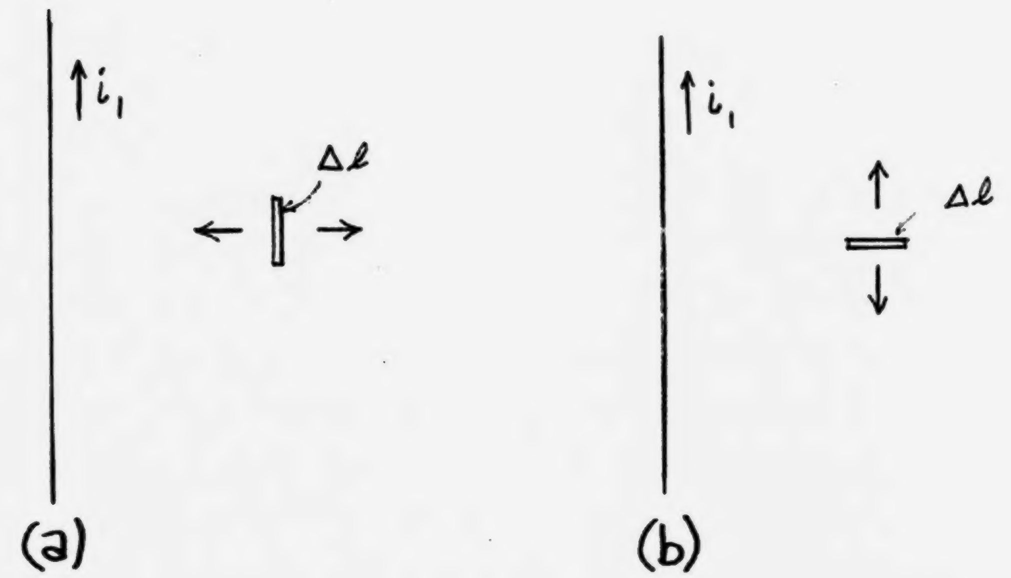

i_1
$\Delta\ell$
(a)
i_1
$\Delta\ell$
(b)

Figure 4-2.

The idea of a *vector product* of two vectors is used to describe the force on $\Delta\ell$. To see how, we shall consider the four conditions illustrated in Figure 4-2, repeated in perspective view in Figure 4-3. The direction of current i_2, flowing in element $\Delta\ell$, is designated by a unit vector $\vec{u}$, and $\vec{F}$ is the force vector. In each case a unit vector $\vec{u}_t$ is shown in perspective view, as a dashed line. This vector is drawn tangentially to a circle centered at wire (1), as indicated in Figure 4-3(a). It will be noted that in each case the right hand screw rule applied in the vector product

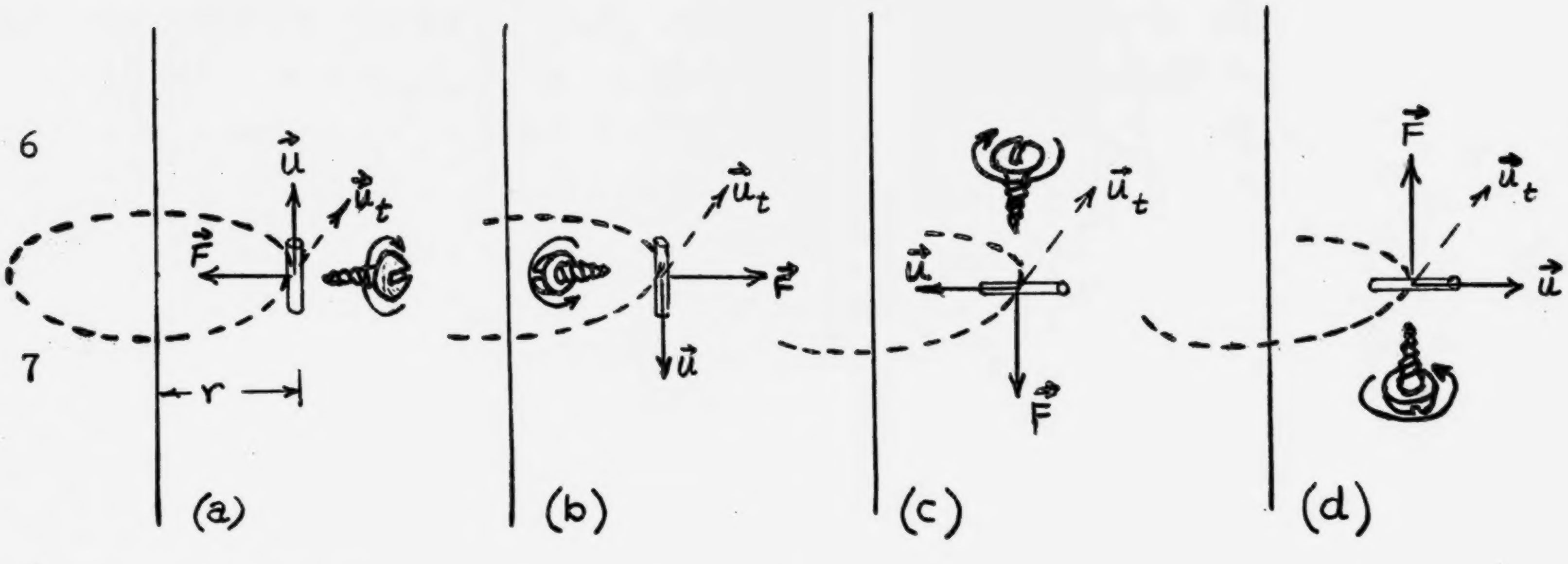

$\vec{u}$
$\vec{u}_t$
$\vec{F}$
r
(a)
$\vec{u}_t$
$\vec{F}$
$\vec{u}$
(b)
$\vec{u}_t$
$\vec{u}$
$\vec{F}$
(c)
$\vec{F}$
$\vec{u}_t$
$\vec{u}$
(d)

Figure 4-3.

$$\vec{u} \times \vec{u}_t$$

will yield the direction of $\vec{F}$. Thus, a unique direction (that of $\vec{u}_t$) can be found, from which the direction of $\vec{F}$ can be derived, when the current direction (defined by $\vec{u}$) is known.

Regarding magnitude of the force, careful analysis of the mechanics involved will show that the magnitude is the same in all situations shown in Figure 4-3, provided that the perpendicular distance r from wire (1) to element $\Delta\ell$ is the same.

Since the force is proportional to $i_1 i_2$, this distance factor must appear explicitly in the equation for force. Also, the factor 1/r must appear, to account for the inverse relationship with distance obtained for Figure 4-1(a). Finally, we should expect force to be proportional to $\Delta\ell$. Thus, experimental results are satisfied by the relation

$$\vec{\Delta F} = k \frac{i_1 i_2 \Delta\ell}{r} (\vec{u} \times \vec{u}_t) \qquad (4\text{-}1)$$

where k is a proportionality factor chosen to take care of units. This constant is usually written $k = \mu_0/2\pi$, where μ_0 is another constant called the permeability of free space. When force, current, and distance are in MKSC units (respectively newtons, amperes, meters), the value of μ_0 is

$$\mu_0 = 4\pi \times 10^{-7}$$

and so we get 2×10^{-7} for the value of k.*

---

*The appearance of the simple rational value 2×10^{-7} for k in Eq. (4-1) may be surprising, in view of the fact that Eq. (4-1) is presented as an experimental law. This is not an accident; it stems from the choice of ϵ_0 which establishes indirectly the unit of charge at such a value that the coefficient is exactly 2×10^{-7}. This procedure of using a non-rational factor (ϵ_0) to define charge, so that the factor comes out rational for B may seem to be somewhat puzzling; it would perhaps seem more logical to choose ϵ_0 as a rational number, since electrostatic phenomena are simpler. The answer lies in the historical development of systems of unit, and the fact that early measurements establishing the ampere were made by measurement of magnetic forces. Magnetic forces are easier to measure in the laboratory, although they are more difficult to analyze theoretically. In the MKSC system of units, adopted as the international standard in 1938, the theory was adjusted to agree with existing practical units wherever possible. Since current and charge are very fundamental, their units were retained.

Now we introduce this value of k, and rearrange to give

$$\vec{\Delta F} = i_2 \Delta \ell \left[\vec{u} \times \left(\frac{2 \times 10^{-7}\, i_1}{r} \vec{u}_t \right) \right]$$

Next, we introduce a symbol (B) to replace the quantity in parentheses and arrive at

$$\vec{\Delta F} = i_2 \Delta \ell \, (\vec{u} \times B\vec{u}_t)$$

where B is

$$B = \frac{2 \times 10^{-7} i_1}{r} = \frac{\mu_0 i_1}{4\pi r} \qquad (4\text{-}2)$$

The quantity $\vec{B} = B\vec{u}_t$ is a vector in the direction of the unit vector $\vec{u}_t$ (or opposite if B is negative). It is called the flux density vector or, alternately, the induction vector. In the MKSC system of units, B is in webers per square meter.* Equation (4-2) is an expression of the Biot-Savart law, giving the flux density, at all points in space, due to a long current-carrying conductor. Thus, using the vector $\vec{B}$, the force equation is

$$\vec{\Delta F} = i_2 \Delta \ell \, (\vec{u} \times \vec{B}) \qquad (4\text{-}3)$$

The validity of this expression has not been completely established by the experiments described, because only two orientations of the $\Delta \ell$ elements have been considered. However, by conducting further experiments, it is found that Eq. (4-3) is indeed valid regardless of the orientation of $\Delta \ell$ with respect to $\vec{B}$. Interpretation of the vector product shows that the magnitude of the force is

$$\left| \vec{\Delta F} \right| = \Delta \ell \left| i_2 \right| \left| \vec{B} \right| \sin \theta$$

---

*In the older electromagnetic system of units, B is in gauss (1 gauss = 10^{-4} webers/sq.m.). In that system, if i_1 is in amperes, r is in cm., then $\mu_0 = 0.4\pi$ and in Eq. (4-3) force is in dynes.

where θ is the angle between the B vector and the conductor, and the force is normal to the plane of B and the conductor.

Plots of $\vec{B}$ at different distances from the wire yield a picture like Figure 4-4(a). At increasing distances from the wire, the arrows become shorter, illustrating, by their length, the 1/r relationship. The state of space in which forces are exerted on current elements (that is, regions in which $\vec{B}$ is not zero) is called a magnetic field. Figure 4-4(a) is a partial plot of the magnetic field surrounding a straight wire. An alternate method of plotting a field is shown in Figure 4-4(b), in which only the directions are indicated by "flow lines".

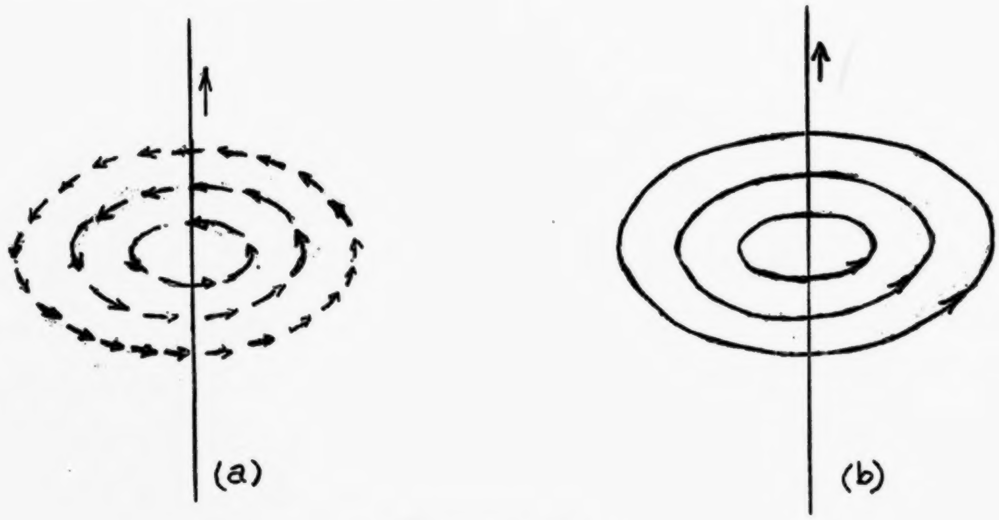


Figure 4-4.

When the force on a long conductor (as distinct from an incremental length) carrying a current i_2 is to be found, incremental forces as given by Eq. (4-3) can be summed to give the total force. Of course, this becomes the integral

$$\vec{F} = i_2 \int_{\text{conductor}} (\vec{u} \times \vec{B})\,d\ell$$

where $\vec{u}$ is a unit vector tangent to the conductor, in the direction of i_2, at each point on the curve. The meaning of this integral can be illustrated by the example in Figure 4-5. The central dot represents the long conductor, and PQ is one side of a square loop. We are to find the force on this piece of conductor. In this case, $d\ell = dy$,

$$B = \frac{\mu_0 i_2}{4\pi\sqrt{b^2 + y^2}} \quad \text{and} \quad \sin\theta = \frac{y}{\sqrt{b^2 + y^2}}$$

giving

$$F = \frac{\mu_0 i_2}{2\pi} \int_0^h \frac{y\,dy}{b^2 + y^2} = \frac{\mu_0 i_2}{4\pi} \ln\left(\frac{b^2 + h^2}{b^2}\right)$$

This force is directed out of the paper, in accordance with the right-hand screw rule as applied to the vector product $\vec{u} \times \vec{B}$.

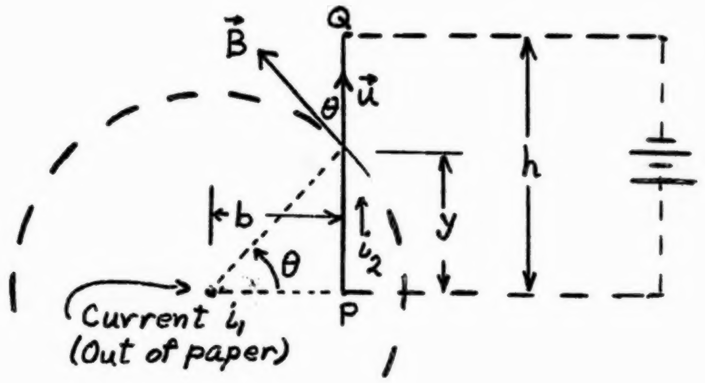
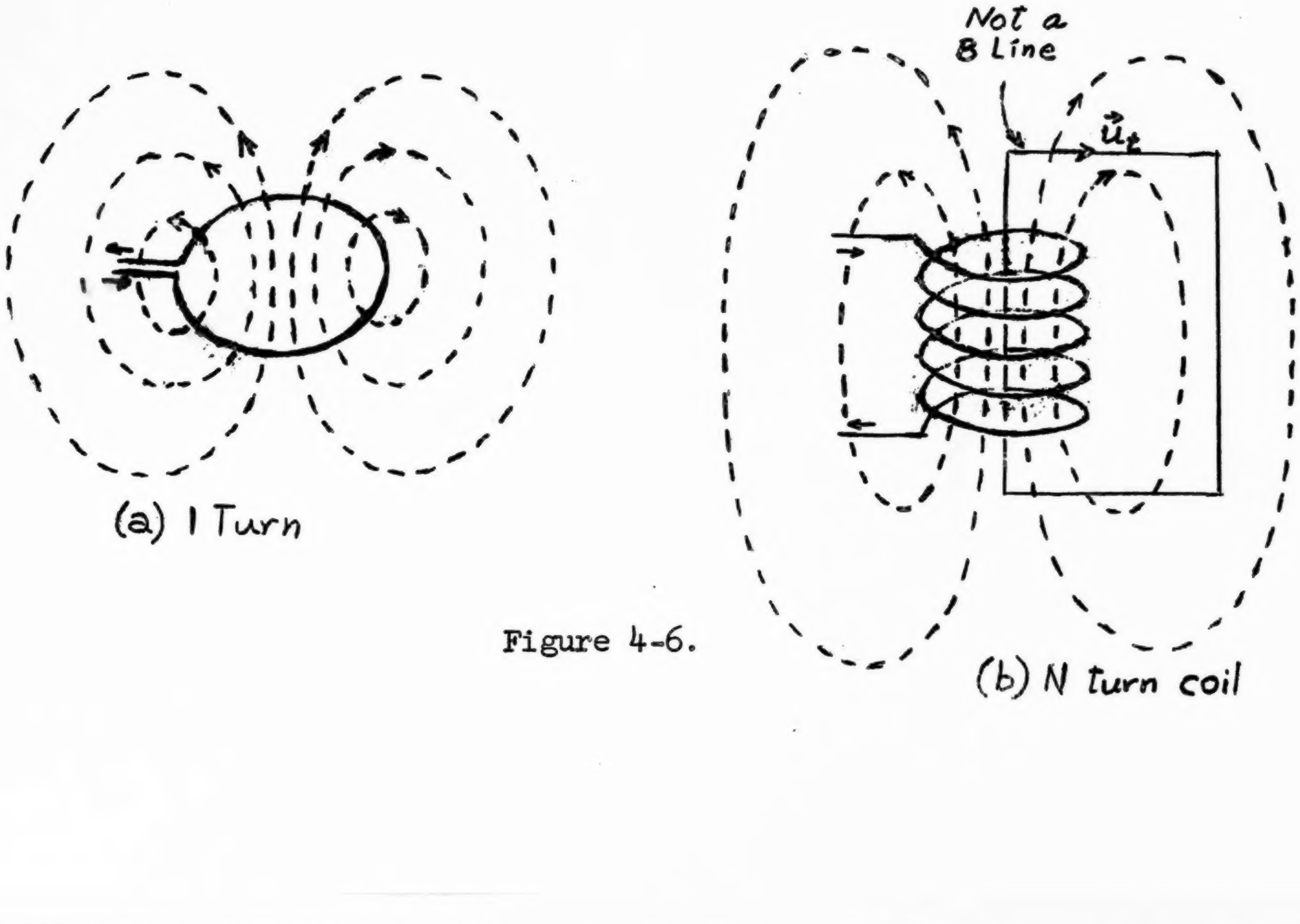


Figure 4-5.

The example chosen to introduce $\vec{B}$, the straight wire, is particularly simple. Problems of finding $\vec{B}$ in the space surrounding other wire shapes are generally more difficult, and are among the topics covered in the subject of field theory. Some illustrations of the structures of the magnetic fields around coils of wire are shown in Figure 4-6.



(a) 1 Turn

(b) N turn coil

Figure 4-6.

One principle can be given in relatively simple terms here, however, which provides some idea of the relationship between B and the current which causes it. For this, we return to the relationship

$$B = \frac{\mu_0 i_1}{2\pi r}$$

for the straight wire. If we take any circle of radius r and multiply B on that circle by the circumference, we get $\mu_0 i_1$. But this product (B times circumference) is a special case of the more general formula

$$\oint B \, d\ell = \mu_0 i_1 \qquad (4\text{-}5)$$

Increments $d\ell$ are along a closed B line (the O on the integral sign merely symbolizes the closed curve). The essential observation to be made in Eq. (4-5) is that i_1 is the current encircled by the closed curve. Furthermore, Eq. (4-5) is a special case of a more general one in which the curve along which the integral is taken is not necessarily identical with a B line. In that case B and $d\ell$ will not necessarily be in the same direction (see Figure 4-7), and so to agree with Eq. (4-5), the component $\vec{B}\cdot\vec{u}_t$ must be used, giving

$$\oint \vec{B}\cdot\vec{u}_t \, d\ell = \mu_0 i_1 \qquad (4\text{-}6)$$

Up to this point, we have only shown that Eq. (4-5) is a special case of Eq. (4-6) and also that the relation $2\pi r B = \mu_0 i_1$ is in turn a special case of

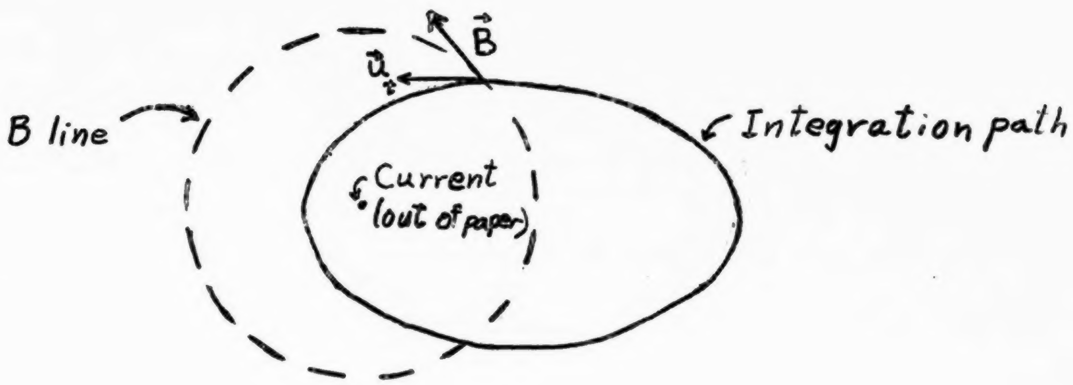


Figure 4-7.

Eq. (4-5). One further generalization has to do with the possibility of having a coil of N turns, like Figure 4-6(b). If the closed curve encircles all turns, the current encircled is Ni_1. Thus, using the solid line in Figure 4-6(b), we have

$$\oint \vec{B}\cdot\vec{u}_t \, d\ell = \mu_0 N i_1 \qquad (4\text{-}7)$$

as an equation which reduces to Eq. (4-6) when N = 1.

In the absence of direct proof for the general case represented by Eq. (4-7), this equation will be taken as a basic postulate. By this statement we mean that Eq. (4-7) has not been proved experimentally for all possible coil shapes and sizes, but its consequences are consistent with all observed phenomena. Equation (4-7) is one form of Ampere's circuital law. Relative to Eqs. (4-6) and (4-7), a word of explanation is needed concerning the relationship between the direction of u_t, which is along the curve of integration and the direction of i_1. They are related in accordance with the right-hand screw rule; the reference direction of i_1 is the direction a right-hand screw will advance when rotated around the curve in the direction of u. Figure 4-7 shows this relationship.

Ampere's circuital law permits the solution of another simple problem, involving the toroidal coil shown in Figure 4-8. From the symmetry of the

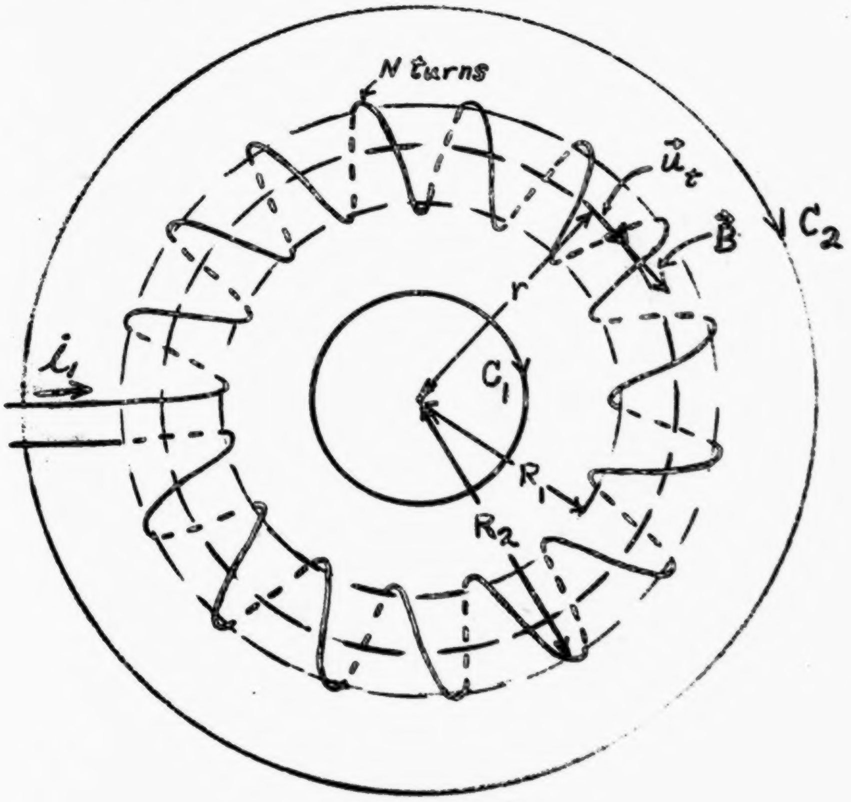


Figure 4-8.

figure, it is evident that, at every point within the coil, B will be tangent to a circle, as indicated. This circle, of radius r, encircles Ni_1 amperes, entering the paper. Thus, the right-hand screw rule shows that u_t should be in the direction shown, and Eq. (4-7) becomes

$$2\pi r B = \mu_0 \, N i_1$$

or $$R_1 < r < R_2 \qquad (4\text{-}8)$$

$$B = \frac{\mu_0 \, Ni}{2\pi r}$$

With the exception of the introduction of N, this is the same as Eq. (4-3). However, Eq. (4-8) applies only for the region inside the coil. For points outside the coil, B is zero. This we can see by observing that any other circles, such as C_1 or C_2, will encircle no current.

4-3. Force on a Moving Charge

The phenomenon of force on a current carrying conductor can be extended to the situation where free charges (as distinct from charges within a conductor) are moving in a magnetic field. A common example is a television picture tube, in which control of the motion of a stream of electrons is partially provided by a magnetic field.

The current in a conductor can be viewed as due to the motion of a row of charges, symbolized by the dots in Figure 4-9. These charges are all of one sign,

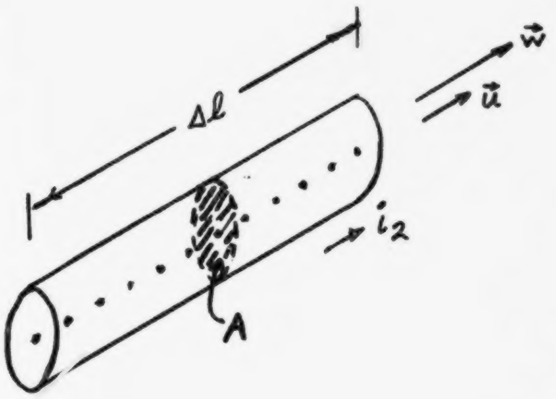


Figure 4-9.

and if the conductor is uncharged, there is an equal number of stationary charges of opposite sign, not shown in the figure. Let $\vec{w} = w\,\vec{u}$ be the average velocity of the moving charges, and let N be the number of them per unit length. Each moving charge has a charge q. In an increment of time Δt each charge will move a distance $w\Delta t$, and therefore in time Δt all charges in this length will pass a given cross section, such as A. Thus, the current is

$$i_2 = \frac{Nq\,w\,\Delta t}{\Delta t} = Nq\,w$$

If the conductor is in a magnetic field, we have seen that the force on it is

$$\Delta\vec{F} = i_2\Delta\ell\,(\vec{u} \times \vec{B})$$

Now we postulate that the force on the wire is actually due to force on the moving charges, and substitute for $i_2\Delta\ell$ in the above, to give

$$\begin{aligned}\Delta\vec{F} &= (Nq\,w\,\Delta\ell)(\vec{u} \times \vec{B})\\ &= (Nq\,\Delta\ell)(\vec{w} \times \vec{B})\end{aligned}$$

since $w\vec{u} = \vec{w}$. Finally, in length $\Delta\ell$ there are $N\Delta\ell$ charges, and so dividing by $N\Delta\ell$ will give

$$\vec{F} = q(\vec{w} \times \vec{B}) \tag{4-9}$$

as the force on one charge. The above postulate implies that this formula will give the force on a moving charge even when it is not in a conductor. Experiment on the deflection of the electron beam of an oscilloscope bears this out, and so Eq. (4-9) should be regarded as giving the force on a moving charge under all circumstances, whether it be a free electron or ion, a physically moving charged object, or the charges within a conductor or semiconductor.

The force described here is known as the Lorentz force.

4-4. Motional Induced Voltage

Having established the idea of a magnetic flux density, and the Lorentz force, we are prepared to consider what happens when a conducting body moves in a magnetic field. Consider the experiment illustrated in Figure 4-10(a), in which a wire PQ is perpendicular to a pair of conducting rails along which it slides with velocity $\vec{w}$ in a direction parallel to the rails. Furthermore, there is a magnetic flux density B which we consider to be uniform over PQ and perpendicular to the plane of PQ and $\vec{w}$.

The conducting wire contains charges which are forced to have a velocity $\vec{w}$ because they must move with the wire. They therefore experience a Lorentz force which pushes them along the wire. Since the free charges are electrons, they move toward P, in the direction of $-(\vec{w} \times \vec{B})$. As a result, a charge separation takes place, as indicated by the polarity markings in Figure 4-10(a). By virtue of this charge separation, there is a potential difference v, which can be measured by a voltmeter. The charge separation takes place very quickly, and the motion of charge stops when the electrostatic force due to the separated charges is exactly equal to the Lorentz force, as indicated in Figure 4-10(b). Since the conductor is perpendicular to the plane of $\vec{w}$ and $\vec{B}$, the Lorentz force acts along the wire, and the work that it does in moving a unit of charge from P to Q is the force per unit charge times $\Delta\ell$. In the notation $\vec{w} = w\,\vec{u}_w$ and $\vec{B} = B\,\vec{u}_B$, since $\vec{w}$ and $\vec{B}$ are perpendicular to each other, on a positive charge this force, of amount Bw, is directed from P to Q. Thus, the work per unit positive charge is

$$e_{PQ} = Bw\Delta\ell \qquad (4\text{-}10)$$

This quantity (e_{PQ}) is called the electromotive force (frequently abbreviated emf) acting from P to Q.* However, from the energy conservation postulate, the work done by the Lorentz force in moving a unit charge from P to Q must equal the work the electric field caused by the charge separation would do in moving it from Q back to P. But this is the potential difference v (or v_{QP}). Thus, for this situation,

$$v = v_{QP} = e_{PQ} \qquad (4\text{-}11)$$

---

*This terminology is a misnomer, because electromotive "force" is a scalar quantity (energy) whereas force is a vector. This incorrect terminology is a heritage from the past which persists due to long usage.

Either of these is loosely termed induced voltage. In view of the identity of v_{QP} and e_{PQ}, one might naturally wonder why a distinction is made between them. As we shall see later, the reason is that when current flows in the wire, v_{QP} is no longer equal to e_{PQ}. They are the same only on open circuit.

In the above example, the moving conductor, B, and w are mutually perpendicular. The result is an equation for e_{PQ} completely in the scalar quantities B, w, and $\Delta\ell$. It is not difficult to show that in the absence of mutual perpendicularity of B and w, the general formula is

$$e_{PQ} = \vec{u}_c \cdot (\vec{w} \times \vec{B})\Delta\ell \qquad (4\text{-}12)$$

where $\vec{u}_c$ is a unit vector from P to Q along $\Delta\ell$. This result will not be proved here.

In the case of a conductor which is extensive enough so that $\vec{B}$ cannot be considered uniform over its length, and may not necessarily be straight, Eq. (4-12) can be regarded as the emf induced in an incremented element of length $\Delta\ell$. An integration can then be performed with respect to a variable ℓ measured along the conductor, giving

$$e_{PQ} = \int_P^Q \vec{u}_c \cdot (\vec{w} \times \vec{B})d\ell \qquad (4\text{-}13)$$

Reference to Figure 4-10(c) will help to clarify the meaning of this integral. Note that at each point on the curve $\vec{u}_c$ is tangent to the curve and directed toward Q.

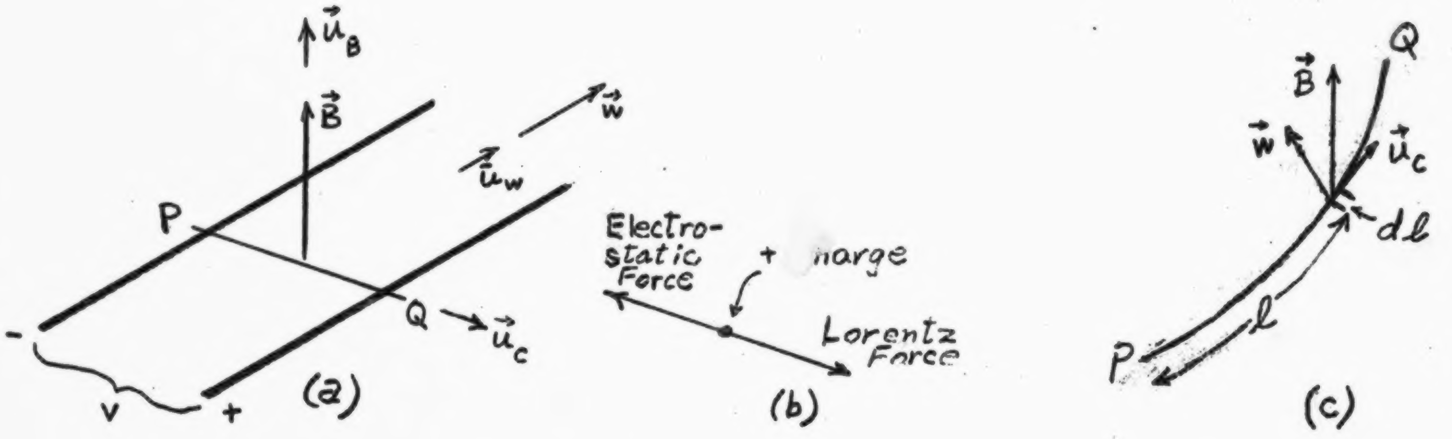


Figure 4-10.

A simple application is shown in Figure 4-11, where PQ is moving with velocity $\vec{w}$, and $\vec{B}$ is due to current i_1 in a long straight conductor. In the region of conductor PQ, $\vec{B}$ is into the paper, as indicated by crosses. It will be found that $\vec{u}_c \cdot (\vec{w} \times \vec{B})$ is negative, and since

$$B = \frac{\mu_0 i_1}{4\pi y}$$

we get

$$e_{PQ} = -\frac{\mu_0 i_1 w}{4\pi} \int_b^{b+h} \frac{dy}{y}$$

$$= -\frac{\mu_0 i_1 w}{4\pi} \ln \left(\frac{b+h}{b}\right)$$

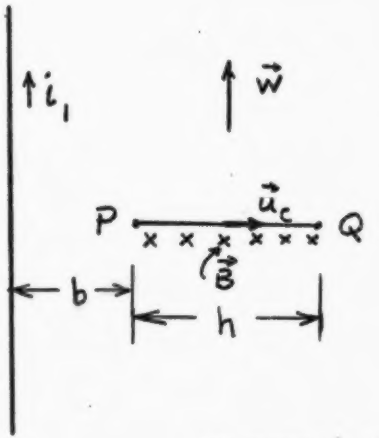


Figure 4-11.

4-5. Faraday Induced Voltage

Next we consider **an** experiment, illustrated in Figure 4-12(a), which involves the effect on one circuit of a changing current in another circuit. A voltmeter is connected to the open ends of a loop of wire, and the loop is near a long wire carrying a current i_1 which varies with time. The wire is in the plane of the loop. Voltage indications on the meter are described by the following observations:

(1) When i_1 is decreasing, the voltmeter will show a negative indication (indicating a potential difference opposite to the meter polarity marks).

(2) When i_1 is increasing, a voltage of the indicated meter polarity will be observed.

(3) The magnitude of the voltage is proportional to the magnitude of the rate of change of current.

(4) For a given rate of change of current, the magnitude of the voltage varies inversely with distance from the wire, and at any given distance is proportional to the loop area.*

Although the above observations involve current i_1, a mathematical relationship describing these phenomena can be obtained by using the magnetic flux density vector ($\vec{B}$), since B is related to i_1 by the Biot-Savart law, $B = \mu_0 i_1/4\pi r$. If $\vec{u}$ is a unit vector directed out of the paper, from the center of the loop, then $\vec{B} = B\vec{u}$, where $B = \mu_0 i_1/4\pi r$. The observed phenomena are accounted for by the formula

$$v_{QP} = -\frac{d(AB)}{dt}$$

where B is the average value of the flux density within the loop, and A is the loop area. The above observations (1), (2), and (3) are accounted for by the negative sign and the fact that B is proportional to i_1; and observation (4) is accounted for by the way B varies with r, and by the inclusion of the factor A.

A further experiment using a coil of N turns of wire will lead to the observation that the voltage is also proportional to N. Thus, for this case, the above equation would include the factor N.

The quantity $\phi = AB$ is called the magnetic flux linking the loop. If the loop had not been normal to the B vector, the induced voltage would be smaller, and investigation would have shown that the component ($\vec{B}\cdot\vec{u}$) of $\vec{B}$ normal to the loop should be used in obtaining ϕ. Thus, a general formula for flux is

$$\phi = A\,\vec{B}\cdot\vec{u} \qquad (4\text{-}14)$$

where $\vec{u}$ is normal to the loop, and $\vec{B}\cdot\vec{u}$ is averaged over the loop surface. This relationship between ϕ and B is, in fact, the origin of the designation of $\vec{B}$ as flux density.

---

* The functional relationships described in (4) are accurately true only for a loop of infinitesimal size.

In similarity with the case of motionally induced voltage, it is convenient to introduce $e_{PQ} = v_{QP}$ as the work done by the changing magnetic field in moving a unit positive charge from P to Q. Thus, using Eq. (4-14) for ϕ, and assuming a loop of N turns, the equation for the electromotive force is

$$e_{PQ} = - N \frac{d\phi}{dt} \tag{4-15}$$

Since, on open circuit, $v_{PQ} = - v_{QP} = - e_{PQ}$, the above equation can be written in terms of potential difference, as

$$v_{PQ} = N \frac{d\phi}{dt} \tag{4-16}$$

It is to be emphasized that Eq. (4-16) gives the terminal voltage only on open circuit.

A word of explanation is needed concerning the sign in this equation. As defined in Eq. (4-14), ϕ is a scalar quantity having a reference direction (like current). This involves a somewhat fictitious view of ϕ as something that "flows", but it is a useful idea. Obviously, the negative sign in Eq. (4-15) has no meaning unless reference directions for scalar quantities e_{PQ} and ϕ are properly defined. The reference for e_{PQ} is indicated by the order of subscripts (i.e., work done in moving positive charge from P to Q). The reference for ϕ is obtained from the order PQ by the right-hand screw rule, by stating that the unit vector $\vec{u}$ in Eq. (4-14) shall point away from area A in the direction a right-hand screw will move when rotated according to a progression from P to Q along the loop. This is illustrated in Figure 4-12(b). The reference direction of ϕ can be regarded as a direction from one side of the loop to the other, in the direction of $\vec{u}$.

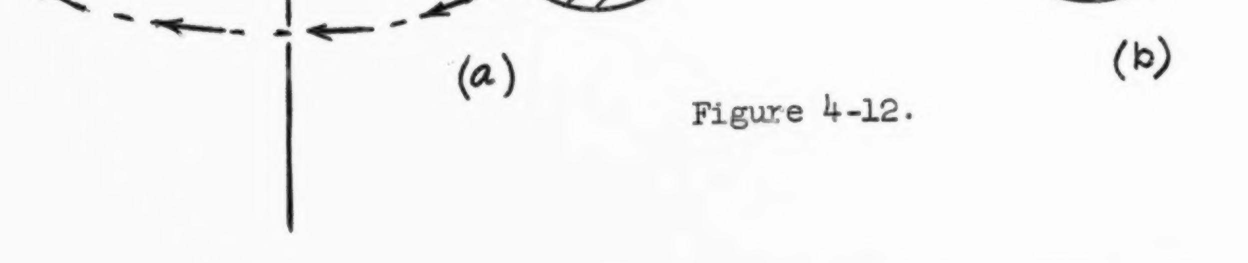


Figure 4-12.

In the above discussions, variation of B within the loop was not considered, having been tacitly accounted for by saying that B in the formulas is the average value. Where B varies over the area, an integral is used to obtain ϕ. This is done by observing that $\Delta\phi = \vec{B}\cdot\vec{u}\Delta a$ is an increment of flux through an area Δa, where $\vec{u}$ is normal to Δa. Thus, in general

$$\phi = \iint_{\text{over surface}} \vec{B}\cdot\vec{u}\,da \qquad (4\text{-}17)$$

should be used for flux whenever B is not constant.

For example, for the rectangular loop in Figure 4-13, the area element

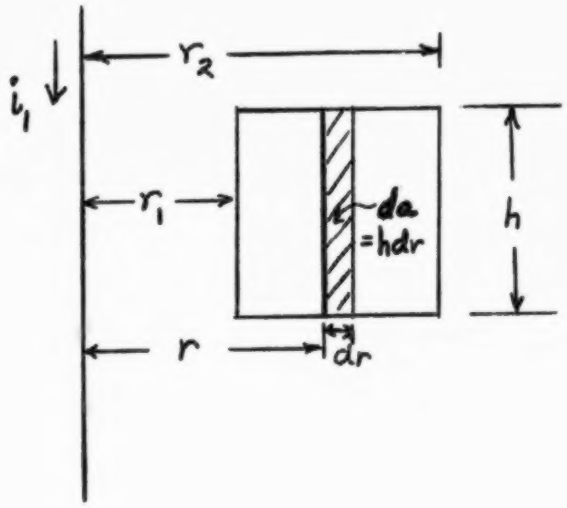


Figure 4-13.

is hdr, and $\vec{B}\cdot\vec{u} = \mu_0 i_1/4\pi r$, giving

$$\phi = \frac{\mu_0 i_1 h}{4\pi}\int_{r_1}^{r_2}\frac{dr}{r} = \frac{\mu_0 i_1 h}{4\pi}\ln\frac{r_2}{r_1}$$

The unit of ϕ in the MKSC system of units is the weber*

---

*In the electromagnetic system, ϕ is in maxwells (gauss × area in cm.^{2}). One maxwell = 10^{-8} weber.

Equation (4-15) or its equivalent, Eq. (4-16), is an expression of the Faraday law of induced voltage. It is presented here for a simple geometrical arrangement, but it is postulated as a generally valid law, regardless of the source of ϕ or the cause of its variation with time. In this example, ϕ varies with time by virtue of the time variation of current i_1. However, ϕ may change with time due to the motion of a fixed current, due to motion of a permanent magnet, or simply due to the motion of a piece of iron in the field.

In some applications a loop moves in a changing magnetic field. In such a case care must be exercised to be sure that Eqs. (4-13) and (4-15) are used correctly. Carelessness in this respect once lead to an interesting historical mistake.*

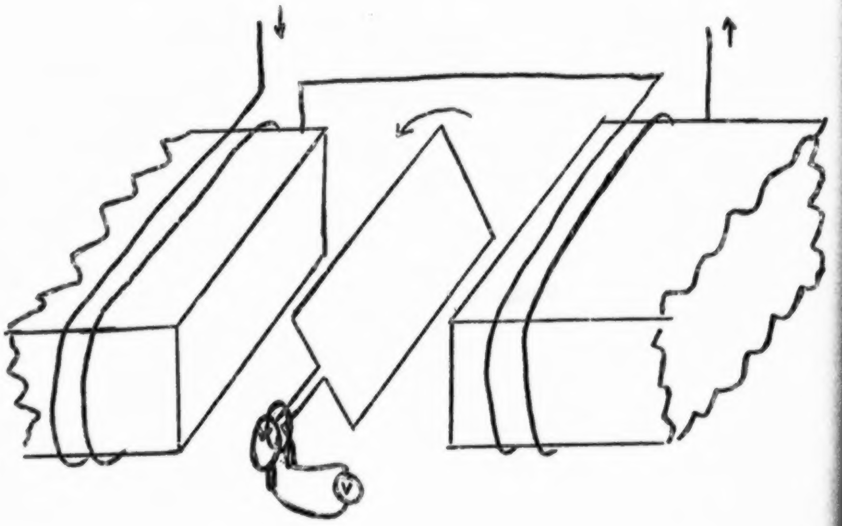
4-6. Elementary Energy Relations

The previous two sections have shown how induced voltage (emf) can be derived from $\vec{B}$ or ϕ, or both, and this was related to potential difference when the device in question was on open circuit. That is, no current was allowed to flow, except for the slight momentary current flow during the short period of time while charge collects on the terminals.

Now consider Figure 4-14 in which the loop between P and Q is subject to

---

*It was once proposed that a commutatorless d-c generator could be made in the manner shown in the figure. It was proposed that the stationary coils should be supplied by alternating current, arranged to reverse once each half revolution of the rotating coil. The argument went that the voltage induced in each side of the rotating coil would always be in the same direction because as the direction of motion reversed at the top or bottom of its circular path the direction of flux would also reverse. However, when the effect of the time variation of flux is included, it is found that d-c is not obtained; and this is in agreement with experiment.



an induced voltage, either by virtue of conductor motion or time varying flux.

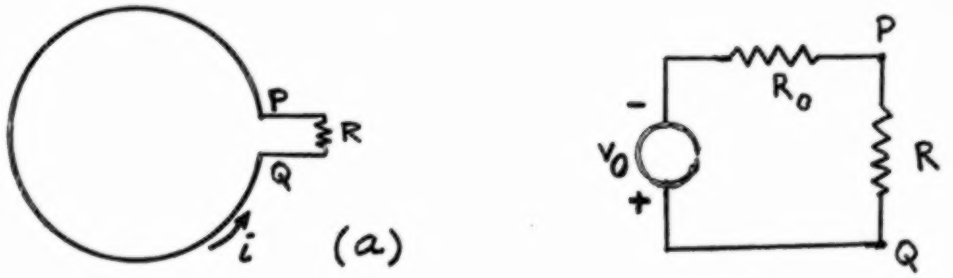


Figure 4-14.

We have seen how to determine e_{PQ}, and it has been recalled that this quantity is the work done per unit charge by the magnetic field in moving this charge from P to Q around the loop. In the present case, the terminals are connected to a resistor R and so a current will flow and the emf e_{PQ} will continually do work. Since e_{PQ} is work per unit charge, and i is units of charge per second, this work will be at the rate $e_{PQ}i$. Using the conservation of energy postulate, this energy rate will equal the sum of $i^2R + i^2R_0$, where R_0 is the resistance of the loop. Thus,

$$e_{PQ}i = i^2(R + R_0)$$

or

$$e_{PQ} = i(R + R_0)$$

But, $i\,R = v_{QP}$, the potential of Q with respect to P. Thus,

$$v_{QP} = e_{PQ} - iR_0 \qquad (4\text{-}18)$$

Note that if $i = 0$, this reduces to the previously obtained equation $v_{QP} = e_{PQ}$.

Equation (4-18) shows the reason for making a distinction between potential difference and emf. Whereas the term "induced voltage" may be used to designate an open circuit potential difference, v_{QP} in Eq. (4-18) should not be called an induced voltage. An emf can be measured by a voltmeter on open circuit (neglecting voltmeter current); or by applying a short circuit to make $v_{QP} = 0$ so that $e_{PQ} = iR_0$, which determines e_{PQ} if i and R_0 are known.

Equation (4-18) is the general relationship between terminal potential difference and current for any device in which there is an electromagnetically induced voltage. If we define the quantity $v_0 = e_{PQ}$ (where v_0 is the open circuit value of v_{QP}) and use it as a voltage source in the circuit of Figure 4-14(b), we see that Eq. (4-18) will be satisfied by this circuit. Thus, this is an equivalent circuit for the source device. It looks like a Thevenin equivalent, but is derived here from different concepts.

When an induced voltage causes a current to flow, as in Figure 4-14(a), a check of the current direction will show that it will always be in such a direction that the magnetic field caused by that current will act to oppose the action creating the emf. This principle is known as Lenz's law. If the emf is due to motion, as in Figure 4-10(a), the force on the conductor due to the current flowing in the B field (the same field which causes the emf) will oppose the motion. Thus, due to this opposition, mechanical work must be done to maintain velocity w. In the case of an emf due to a changing flux, as in Figure 4-12(a), the current caused by this emf when the circuit is closed will create a flux opposing the flux change which causes the emf.

Lenz's law is a consequence of conservation of energy, being a statement that there is an opposing action to any action which will cause a current to flow, such opposing action necessitating an expenditure of energy on the circuit.

4-7. Ferromagnetic Material

The next experiment to be considered is illustrated in Figure 4-15, where $\vec{B}$

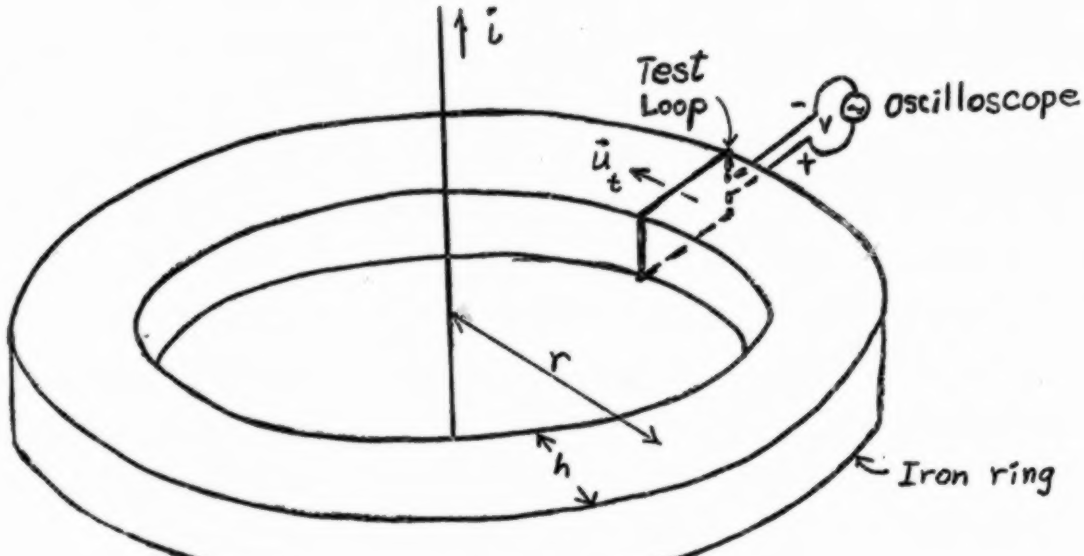


Figure 4-15.

is to be found within the circular iron ring encircling a long straight current-carrying conductor, as shown in the figure. The cross-sectional area of the ring is uniform, and dimension h is very small compared with r.* Obviously, the force on a test conductor cannot be measured within the iron, and so a different method of measuring B must be used.

The total flux through a cross section of the iron can be determined by placing a loop of conducting wire in the manner shown and connecting this to a voltage recording device such as an oscilloscope. Although such an arrangement cannot measure ϕ directly, it can be used to measure a change in ϕ. Thus, suppose i is increased in a sequence of jumps, in the matter illustrated graphically in Figure 4-16(a). With each increase in current, the flux will increase an amount $\Delta\phi$. Now, since

$$v = \frac{d\phi}{dt}$$

at any one of the jumps, say the second, we get

$$\Delta\phi_2 = \int_{t_1}^{t_1+\Delta t} v\,dt \qquad (4\text{-}19)$$

This integral can be evaluated from the oscilloscopic record of v as a function of time.** By making a succession of such measurements, the data points on the solid curve in Figure 4-16(b) are obtained.*** It is found that as the current increases by equal increments, successive increments of ϕ become smaller, causing the curve to flatten out. This phenomenon is called saturation.

---

*The reason for this requirement is that in view of Eq. (4-2) we should expect B to vary with radius. By restricting h to a small value, there cannot be muc variation of r to all points in the ring, and thus B will be nearly constant over a cross section.

**This integral is normally evaluated by the use of a ballistic galvanometer, but in the interest of maintaining a direct attack on the immediate problem, w avoid this detail.

***Readers with awareness of the phenomenon of hysteresis will realize that this curve will be obtained only if the iron is initially unmagnetized. This state can be attained by starting with a large value of i, reversing i to a slightly smaller value, and repeating this process until the current is zero.

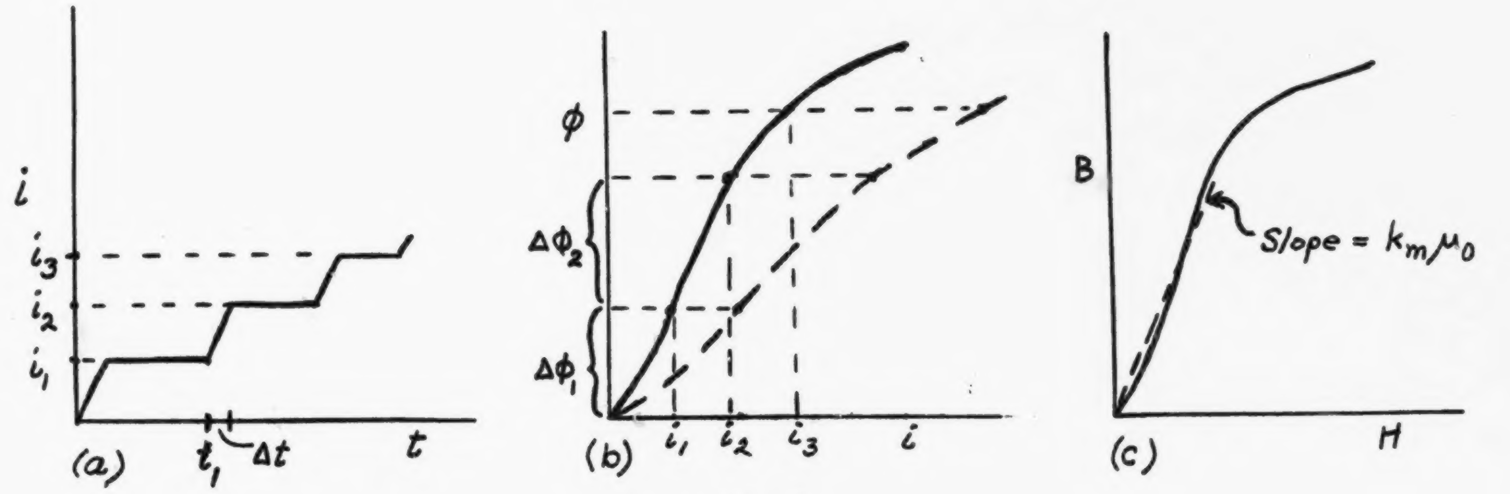


Figure 4-16.

The curve of Figure 4-16(b) is a property of a particular ring. In order to attain a curve which yields intrinsic properties of the magnetic material, independently of physical dimensions, a further experiment is needed, using a ring of different diameter, say twice as large. A new curve will then be obtained, like the one shown by the dashed line in Figure 4-16(b). The amount of current i required to yield a specific value of ϕ will be doubled. Thus, if we had plotted $i/2\pi r$ instead of i as the abscissa, the *same* curve would have been obtained in both cases. Thus, by using $i/2\pi r$ as the abscissa, the diameter (or circumference) parameter is eliminated. Furthermore, the cross-sectional area A can be eliminated by dividing ϕ by A, to give B.⊕ The result will be a similar curve with changed labels on the axes, as shown in Figure 4-16(c). A new variable (H) has been introduced such that

$$H = \frac{i}{2\pi r}\ , \quad \vec{H} = H u_t \qquad (4\text{-}20)$$

Since dimensions of the ring have been eliminated from consideration in Figure 4-16(c), the B-H curve is a property of the material, called a *normal magnetization* curve.

$\vec{H}$ is a fundamental magnetic quantity, called the *field intensity vector*. In the MKSC system of units, H is expressed in amperes per meter.

⊕This neglects the slight variation of B over the cross section.

Let the formula

$$B = f_m(H) \tag{4-21}$$

describe the functional relationship between B and H, such as the graph in Figure 4-16(c). In iron and in similar media, which are called ferromagnetic, the relationship is non-linear. In air, the relationship can be determined by comparing Eqs. (4-2) and (4-20); it is seen to be

$$B = \mu_0 H \tag{4-22}$$

This is a linear relationship showing that B and H are proportional.

There are other materials than air in which B is proportional to H. For these, we write

$$B = k_m \mu_0 H \tag{4-23}$$

where k_m is a constant called the relative permeability. If $k_m < 1$, the material is called diagmagnetic; if $k > 1$, it is called paramagnetic. However, the value of k_m for such materials is but little different from 1; for example, one of the larger values is 1.003 for liquid oxygen.

The linear relationship in Eq. (4-23) is sometimes used, even for ferromagnetic materials, as a linear approximation for the general relationship in Eq. (4-21), as indicated by the dashed straight line in Figure 4-16(c). This is often done in preliminary design calculations, where the advantages of a linear equation (as distinct from a graph) outweigh the inaccuracies involved.

Examples of normal magnetization curves for several ferromagnetic materials are shown in Figure 4-17. Also,included is the straight line $B = \mu_0 H$. It is apparent that B can be increased very appreciably by the introduction of iron, an important consideration in the construction of many items of electrical apparatus.

In most cases (and all cases to be considered in this text) $\vec{H}$ is a vector in the direction of $\vec{B}$, where the relationship between their magnitudes is either given graphically or by Eqs. (4-22) or (4-23). The exception occurs in certain materials which are non-isotropic (materials which exhibit different

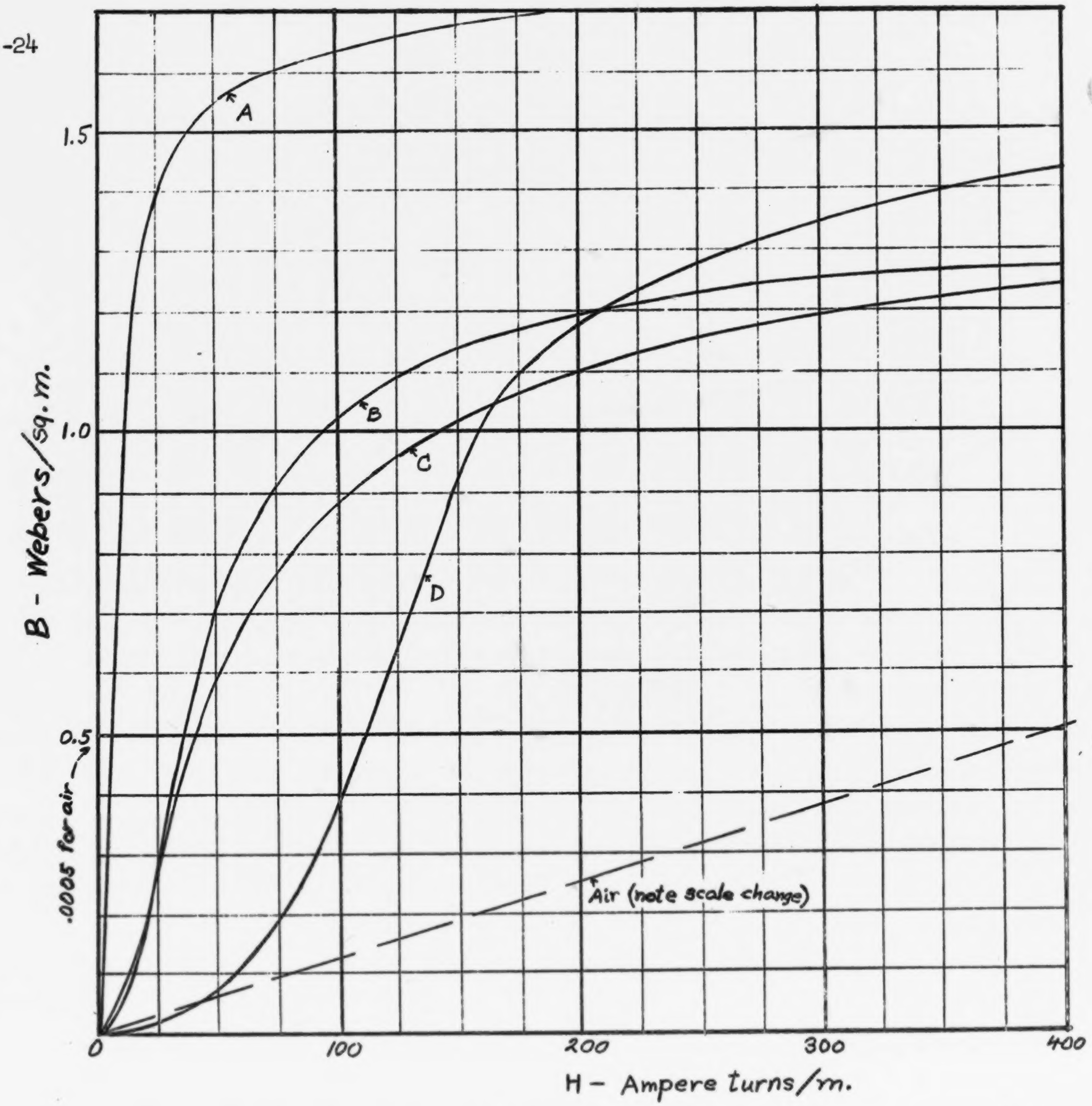


Figure 4-17. Normal magnetization curves (from U.S. Steel Handbook)

Legend

- A Grain oriented Transformer "66" } Silicon Steels
- B Radio Transformer "58" } Silicon Steels
- C Transformer "52" } Silicon Steels
- D Pure iron (annealed)

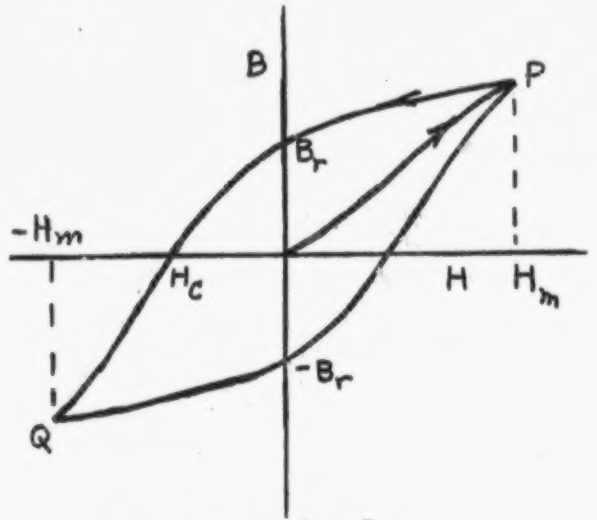


Figure 4-18.

properties in different directions), in which $\vec{B}$ and $\vec{H}$ do not necessarily have the same direction.

4-8. Hysteresis

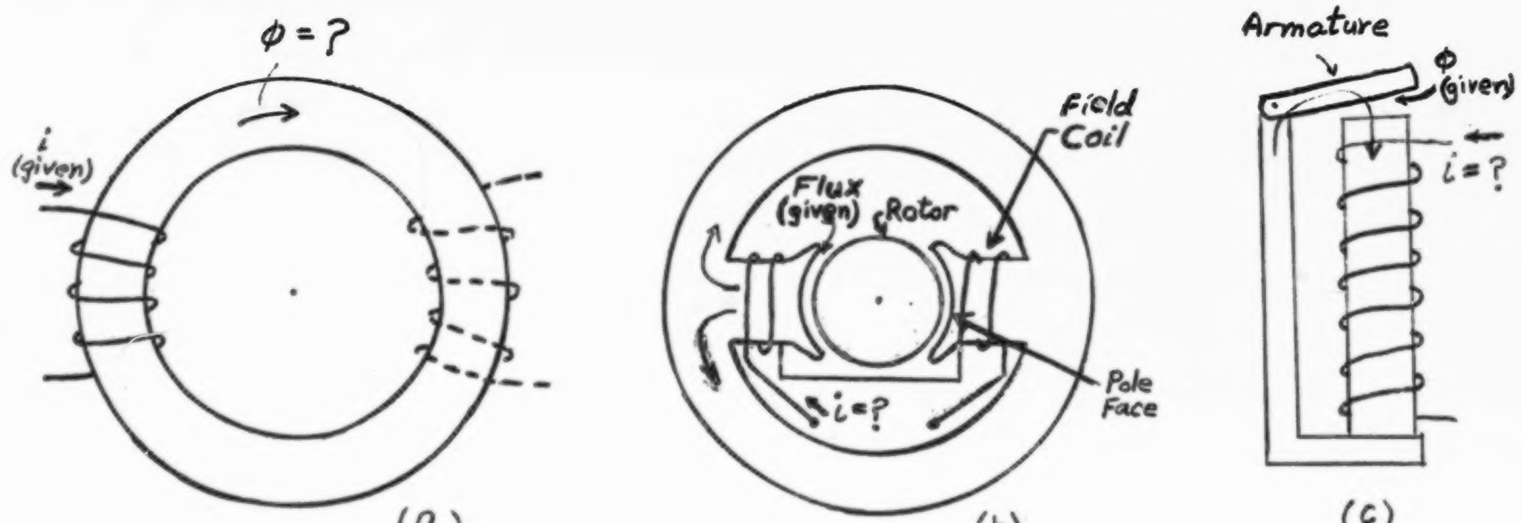
Again referring to Figure 4-15, assume the current has been increased, bringing the magnetic state to point P in Figure 4-18 and then let the current be decreased in steps. With each jump downward, $\Delta\phi$ can again be obtained from the integral of voltage. It will be found that demagnetization will take place along a curve higher than the original one, as indicated by the arrow. If the current is then carried to negative values, as far as $-H_m$, and then back to H_m, a closed curve called a hysteresis loop will be obtained.

Ordinate B_r, where the curve crosses the vertical axis represents the flux density remaining in the iron when the current has been reduced to zero, is called the remnant flux density, and abscissa H_c, the value of H required to reduce B to zero is called the coercive force. The phenomenon of hysteresis described here further complicates computations involving magnetism in iron. Not only is the material non-linear (B not proportional to H) but the relationship between B and H is multivalued. For a given H, the value of B depends on the previous magnetic history of the iron. In many cases approximate calculations are made using the normal magnetization curve. However, there are applications in which the hysteresis phenomenon is of paramount importance. One of these is in devices involving "magnetic memory," such as tape recorders and memory devices used in computers. The memory capability arises from the above-mentioned fact that magnetic state depends on what has happened in the past. For example, by measuring the flux at zero current (giving either $+B_r$ or $-B_r$) it is possible to know whether the magnetization was previously at P or Q.

4-9. Magnetic Circuits

A magnetic circuit is an arrangement of ferromagnetic materials forming closed paths, sometimes including short gaps of air or of other non-magnetic materials, so as to form an easy path for magnetic flux. The source of the flux consists of one or more current-carrying coils of wire surrounding some portion of the path.

Figure 4-19 shows some examples of magnetic circuits, taken from typical electromagnetic devices. The iron ring shown at (a) might be the "core" of a transformer on which there are two or more windings. With a single winding, the device is an inductor. To determine the amount of flux in the iron due to a given current in one of the windings would constitute a magnetic circuit problem. As another example, consider Fig. 4-19(b) which shows the magnetic circuit of a d-c generator or motor. Suppose it is a generator, to operate at a specified speed and voltage. This will determine the required value of B in the airgap between the rotor and the poles, by way Eq. (4-10), and thus a certain airgap flux is known to be required. Final solution of this problem would then determine the number of turns and current on the "field" windings required to produce this flux. This is also a magnetic circuit problem. Our final example, Figure 4-19(c), shows an electromagnetic relay. The force on the moving armature is a function of the airgap flux, and thus a certain flux will be required to cause the relay to close. If this required flux is known, a magnetic circuit problem is then solved to determine the specifications of the coil (current and number of turns.)



(a) (b) (c)

Figure 4-19.

We shall now consider how magnetic circuit problems like the above can be solved. As a start, we return to a consideration of a uniform iron ring, like Figure 4-15, but with a variety of coil arrangements as in the examples in Figure 4-20. The case shown at (2) is the simplest. It is similar to Figure 4-8, except

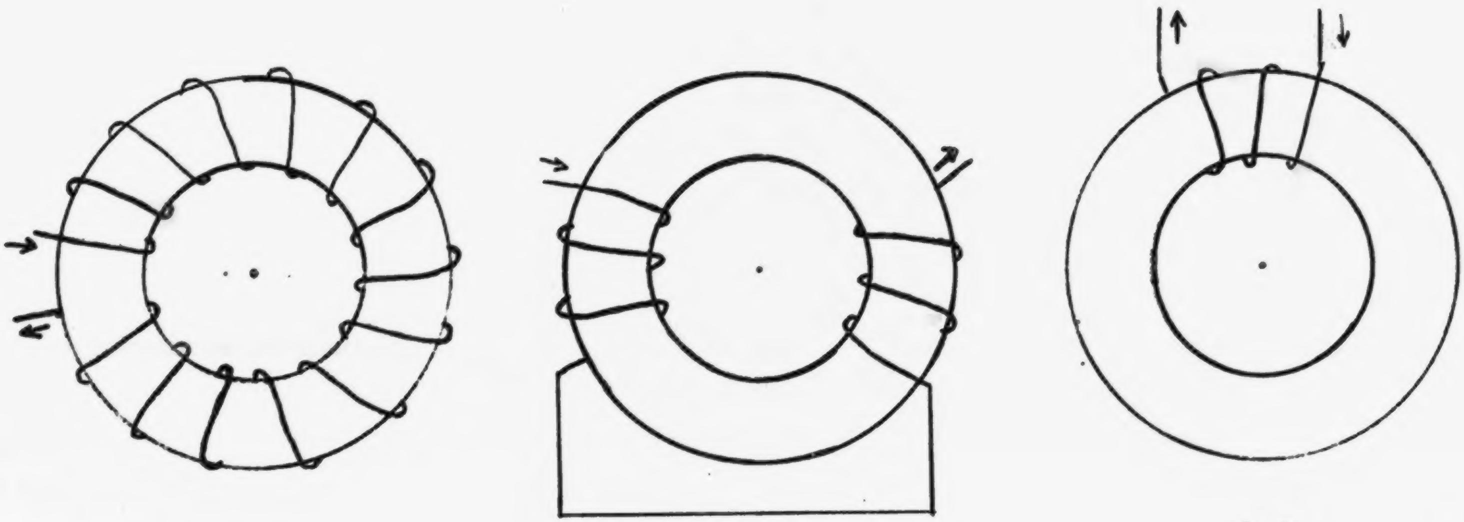

(a) (b) (c)

Figure 4-20.

that now the coil is wound on an iron core. It can be shown experimentally that if the product Ni in Figure 4-20 is the same as i in Figure 4-15, and if the iron ring dimensions are the same, then the flux in the iron will be the same. Now suppose the product Ni is also the same in the examples at (b) and (c) of Figure 4-20. It will be found that the flux in the iron is approximately the same as before, although by moving a test loop (like the test loop of Figure 4-15) to various points around the ring it will be found that there is a slight variation of flux, and that the flux is slightly larger near the windings.

This empirical result is significant, because it implies that the flux in a ring of magnetic material of given dimensions is approximately dependent only on the total current linking the ring (current in the wire times the number of turns, in the case of a coil). It is concluded that B and therefore H are nearly the same in each case, and that Eq. (4-20) approximately applies if we replace i by Ni, giving

$$2\pi r H = Ni \quad \text{(approximately)} \qquad (4\text{-}24)$$

The value of H at a given point (fixed r) depends on the product Ni. The same H is obtained for different values of N and i, so long as the product is the same. Ni is called magnetomotive force,* abbreviated mmf, and is given the symbol U. Its unit is the ampere. (Its unit is sometimes taken to be an ampere-turn, but a turn has no dimensions.)

Let us tentatively leave this approximate relationship and consider the further experiment shown in Figure 4-21. In this case the ring shown at (a) includes an airgap. It is assumed that the airgap length ℓ_a is small compared with the linear dimensions d and h of the cross section. It is also assumed that there is some means for closing the airgap temporarily, say by bending the ring slightly, as shown in Figure 4-21(b). Using a test loop like the one shown in Figure 4-15, two magnetization curves can be determined, one for the ring without the airgap, and the other with the airgap as shown in Figure 4-21(c).

Consider a specific value of flux as indicated by one of the horizontal lines in Figure 4-21(c). When there is no airgap, the required current is determined by U_i, a specific value of Ni obtained from the magnetization curve. When the airgap is included, it is found that the mmf must be increased by an amount U_a in order to maintain the same flux in the iron.

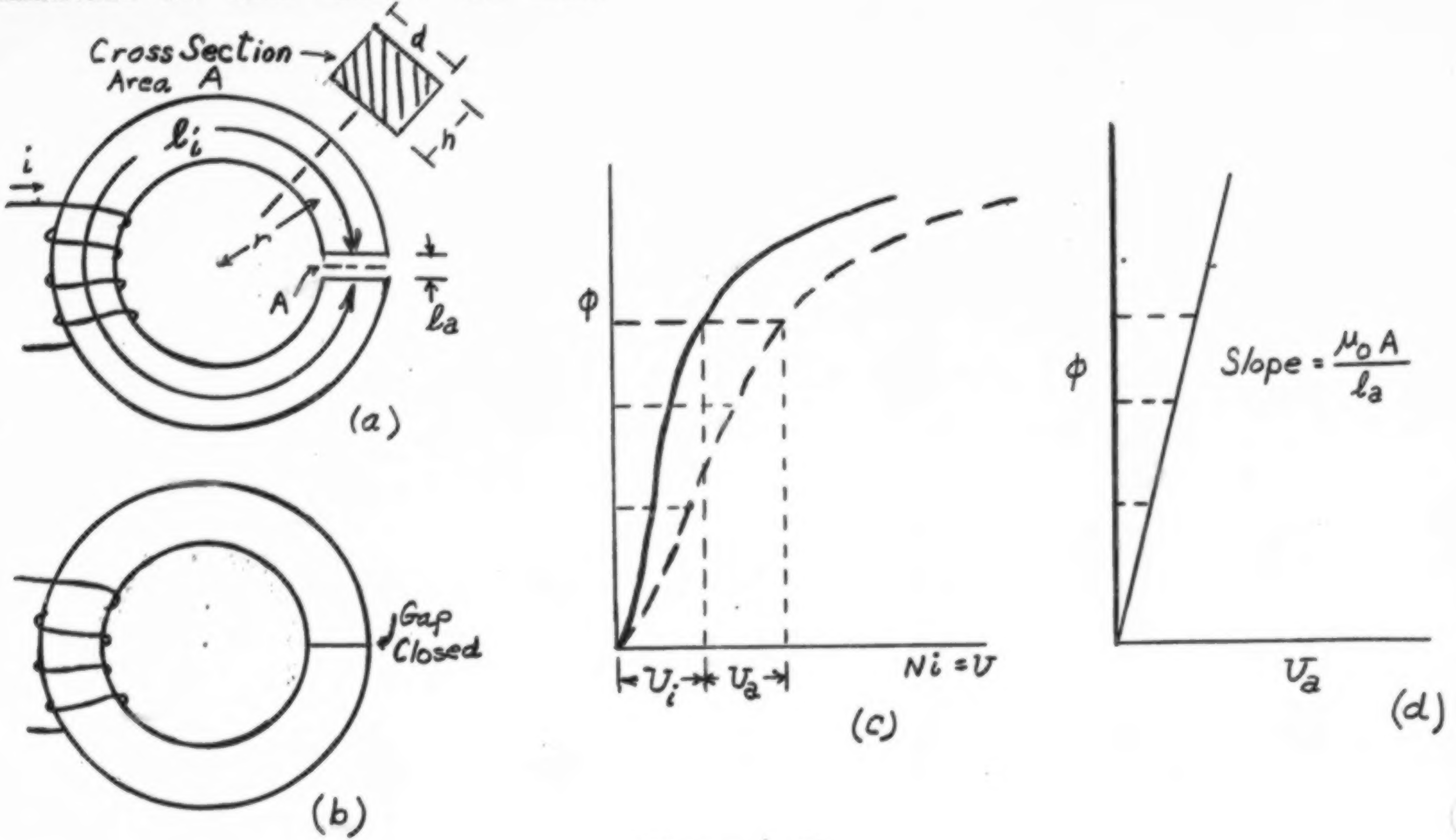


Figure 4-21.

---

*This is another example of incorrect terminology which persists due to long usage. Magnetomotive "force" is not a force; it is a scalar, not a vector.

Recalling that it is an approximation to assume that flux is the same everywhere in the ring, it would be meaningless to retain the slight variation of H with r implied by Eq. (4-24). Instead, we will let ℓ_i be the average circumference of the entire ring when the gap is closed; this is the same as the average length of the iron when there is a short gap. Thus, for the closed-gap situation, Eq. (4-24) becomes

$$H_i \ell_i = U_i \qquad (4\text{-}25)$$

where H_i signifies H in the iron. Since $B = \phi/A$ is the same for both cases, and since B and H_i are related by a magnetization curve which is a property of the iron, H_i is the same in both cases. If the quantity U_a is picked off this curve, for various values of ϕ, and plotted as a function of ϕ, as in Figure 4-21(d), a straight line will be obtained. The slope of such a line can be obtained from experimental measurements, and will be found to be very close numerically to $\mu_0 A/\ell_a$. Thus, the equation of this line is

$$\phi = \frac{\mu_0 A}{\ell_a} U_a \qquad (4\text{-}26)$$

Let us now imagine a flat surface of area A in the airgap, as indicated in Figure 4-21(a), and assume that the flux through this area is ϕ, the same as the flux in the iron. This is clearly an assumption at this point. Dividing both sides of Eq. (4-26) by A, gives B_a on the left, and so we get

$$B_a = \mu_0\left(\frac{U_a}{\ell_a}\right)$$

But this is of the form $B = \mu_0 H$ (the relationship between B and H in air) and so, if the assumption about flux being the same in the iron and the airgap is correct, then

$$U_a = H_a \ell_a \qquad (4\text{-}27)$$

Now, since

$$U_i + U_a = Ni$$

from Figure 4-21(c), it follows that

$$H_i \ell_i + H_a \ell_a = Ni \qquad (4\text{-}28)$$

This is a tentative result, based on an assumption that ϕ is the same on both sides of the gap interface between iron and air.

We shall now show that there is a certain similarity between Eq. (4-28) and Ampere's circuital law as given by Eq. (4-7). Referring to Eq. (4-7), if both sides are divided by μ_0, we get

$$\oint \vec{H} \cdot \vec{u}_t \, d\ell = Ni \qquad (4\text{-}29)$$

According to the original statement of Ampere's circuital law, the integration path for the above equation is wholly in air. Nevertheless, we shall tentatively apply it to the closed path in the iron ring and airgap of Figure 4-21(a). We find that in iron $\vec{H}\cdot\vec{u}_t = H_i$, and in the airgap $\vec{H}\cdot\vec{u}_t = H_a$ (H_i and H_a are both constant) and so the integral on the left of Eq. (4-29) reduces to the left-hand side of Eq. (4-28).

On the clue that Eq. (4-29) applies to a special case having part of the path in iron, we shall now introduce a postulate to the effect that Eq. (4-29) applies to all possible integration paths in all situations. This is the general Ampere's circuital law.

The condition $\phi_a = \phi_i$ used in arriving at the above must also be stated as a general postulate. This is done most easily in terms of the flux entering and leaving a closed surface. Specifically, referring to any hypothetical closed surface, as in Figure 4-22(a), it is postulated that the flux entering the volume equals the flux leaving. This postulate is known as Gauss' law for magnetic fields. To show that this postulate is consistent with $\phi_a = \phi_i$, we refer to Figure 4-22(b), in which a "pill box" shaped volume is used, with a side surface of infinitesimal height, so that any flux leaving this surface will be negligible (if B should have a component normal to this surface). One face of the surface is in the iron, and the other face is in the airgap. Thus, ϕ_i is the flux entering, and ϕ_a is the flux leaving. Gauss' law states they are equal, showing that the original assumption $\phi_a = \phi_i$ is consistent with Gauss' law.

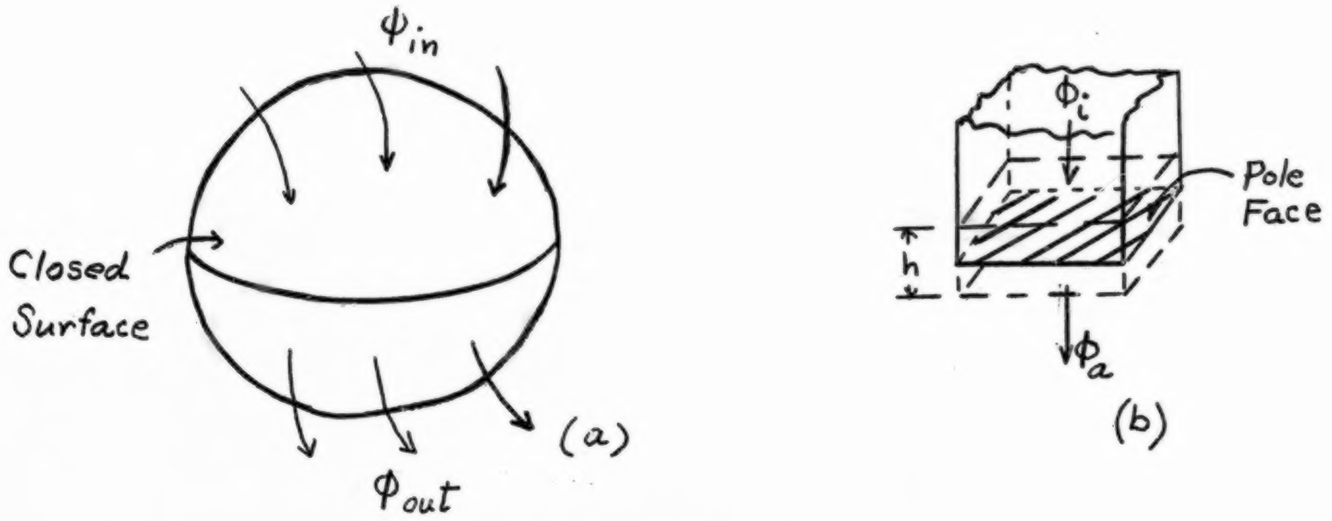


Figure 4-22.

These two postulates are fundamental to the theory of magnetism and have far-reaching consequences which can be checked by experiment. All evidence indicates they are true. We may therefore regard them as fundamental laws, although in reality they remain postulates because their validity is based on evidence now available.*

Before proceeding to a description of methods of solving magnetic circuits, some interesting consequences of the above two laws can be shown, with reference to Figure 4-23(a). A closed integration path a-b-c is shown. Obviously, it does not link a current, and so for this path

$$\oint \vec{H}\cdot\vec{\mu}_t \, d\ell = 0$$

But that part of the integral due to the path in iron and airgap is $2H_i\ell_1 + H_a\ell_a$, where the lengths ℓ_i and ℓ_a are shown in the figure. Thus,

$$\int_{\substack{c \\ \text{(air)}}}^{a} \vec{H}\cdot\vec{u}_t \, d\ell = -(2H_i\ell_1 + H_a\ell_a)$$

The quantity on the right is not zero, and so it is concluded that H along the air path c-a cannot be identically zero. Since $B = \mu_0 H$, B is also not

---

*The "law" of conservation of energy, which was really a postulate, is a good example of the tentativeness of a postulate. Until the discovery of nuclear reactions, this postulate satisfied all available evidence. However, it is known that it is not universally valid, and so can be used only in those situations in which there is no mass-energy interchange. During the 1940s there was a flurry if excitement in scientific circles concerning purported experimental evidence that "magnetic charge" had been discovered. If free magnetic charges exist, then Gauss' law for magnetic fields, as stated here, will not be true. The evidence for free magnetic charge was not conclusive, and so we continue to rely on Gauss' law.

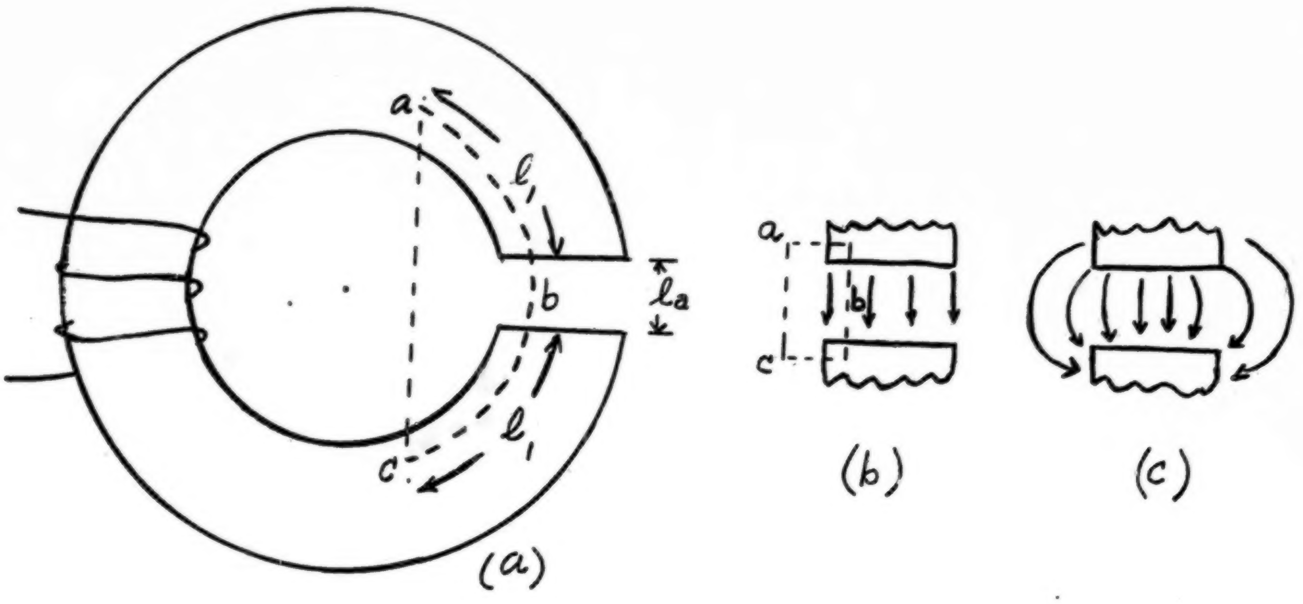


Figure 4-23.

identically zero.* We therefore conclude that there must always be some flux in the air surrounding a magnetic circuit. There is no such thing as a magnetic insulator, and in this sense a magnetic circuit differs from an electric circuit.

Let us pursue the same line of reasoning to investigate the airgap flux. If it is assumed that the field lines go straight across the gap as in Figure 4-23(b), an integration path a-b-c can again be used, showing that B cannot be zero everywhere along c-a, outside the gap. Accordingly, it is necessary for there to be fringing of the gap flux, as shown in Figure 4-23(c), by an amount sufficient to satisfy Ampere's circuital law.

The flux lines surrounding the ring will be somewhat as shown in Figure 4-24(a). Gauss' law gives additional information, if applied to the volume element shown at the top of this figure. If ϕ is the flux in the iron entering the volume at the left, an amount $\Delta\phi$ will leave the iron through the sides. Thus, by Gauss' law, the flux leaving the right, in the iron, is $\phi - \Delta\phi$. Thus, the flux in the iron gets progressively weaker with increasing distance from the coil. In many cases, however, the variation is negligible.

The difference between the flux through the coil, ϕ_t, and the flux ϕ_g at the

---

*We say "not identically zero" meaning not zero at all points. We cannot conclude, however, that there are no points where B and H are zero.

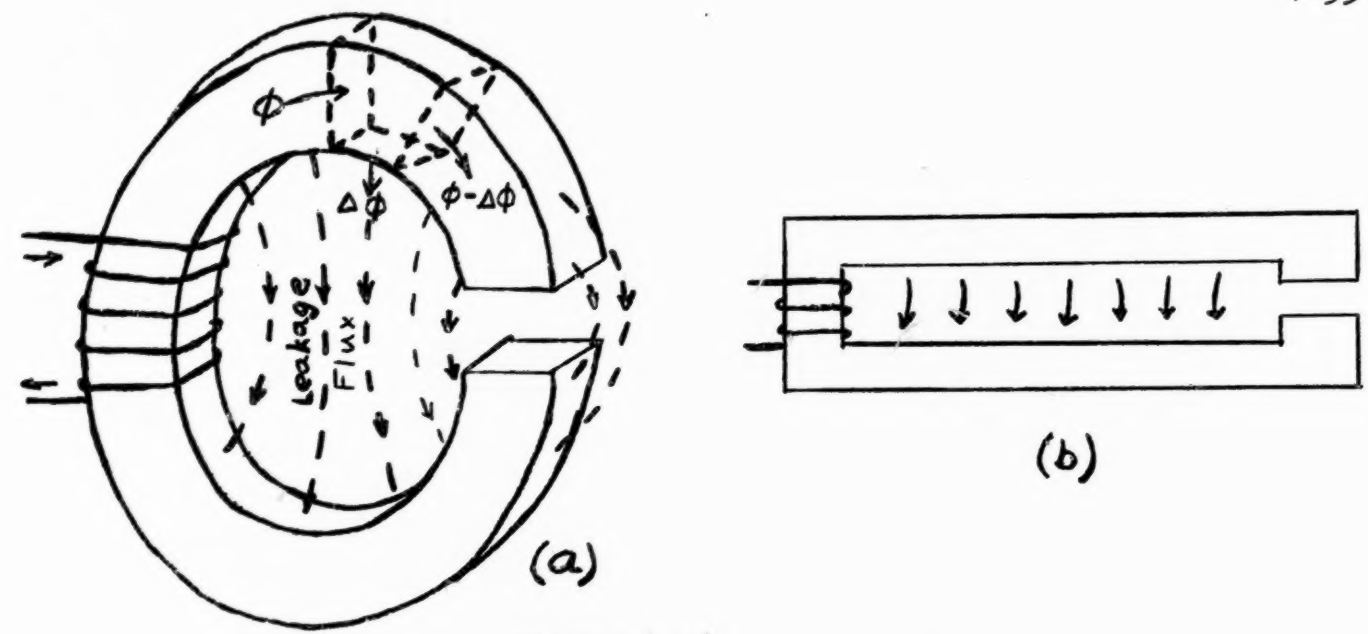


Figure 4-24.

gap, namely

$$\phi_\ell = \phi_t - \phi_g$$

is called the leakage flux. An arrangement like Figure 4-24(a) can be treated a a magnetic circuit only if ϕ_ℓ is small compared with ϕ_g. Whether or not this is true depends on the permeability of the iron, and also on the geometrical shape. For example, an arrangement like Figure 4-24(b) can have a very large leakage flux.*

4-10. Solution of Magnetic Circuits

Reference should again be made to Figure 4-19, as illustrations of typical magnetic circuit problems. Particular note should be made of the fact that airgaps are often essential, as in the case of a motor or generator where mechanica clearance is necessary, or in the relay where opening and closing motion is possible only if there is an airgap. Also, it is to be noted that in the generator

---

*It can be shown that for a reasonably dimensioned circuit (not distorted like Figure 4-24(b)) without airgap the ratio of core flux to leakage flux is approximately

$$\frac{\phi_c}{\phi_\ell} = k_m\left(\frac{\ell_c}{\ell_i}\right)$$

where ℓ_i and ℓ_c are respectively the lengths of the iron path and the coil length, and k_m is the relative permeability of the iron. When there is an airgap, this ratio is smaller in proportion to the reduction in ϕ_c due to the airgap.

there are two magnetic circuits which share a common branch through the center.

The previous section established the principles upon which solution of magnetic circuits is based, including the notion that such solutions are always approximate. We reiterate these principles here, in terms of the example in Figure 4-25.

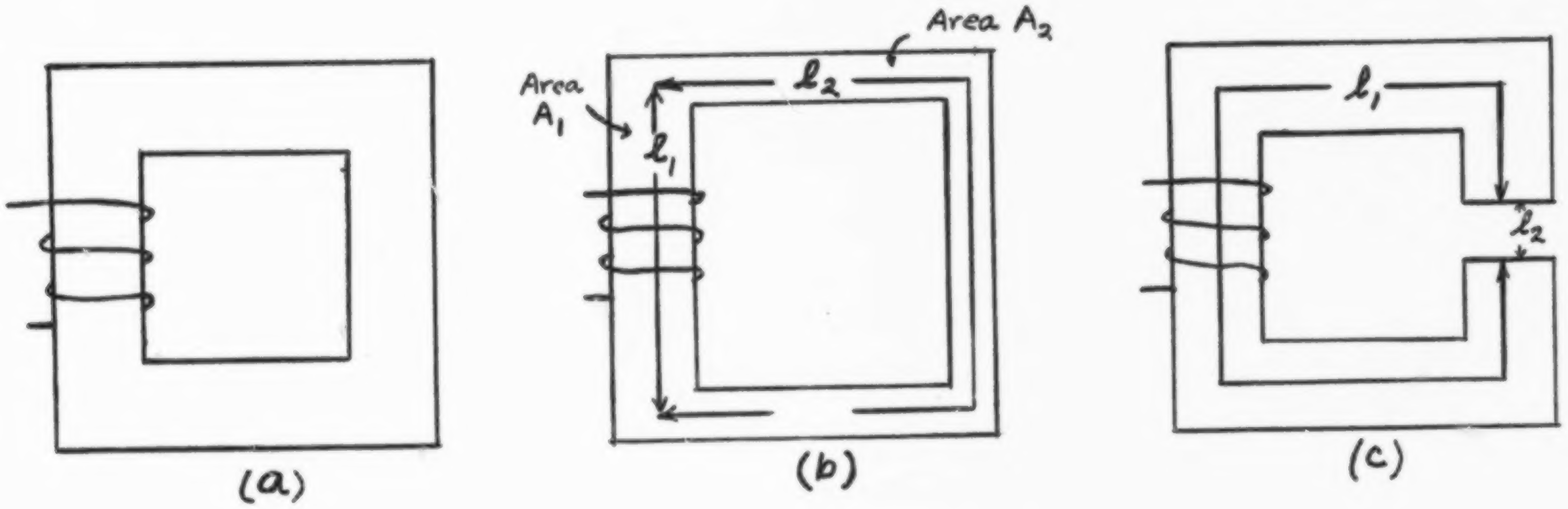


Figure 4-25.

(1) Branches are identified such that in each branch the cross-sectional area is the same, the material is the same, and the flux is the same. Thus, in example (b), two branches are identified and labeled according to their lengths; ℓ_1, ℓ_2.

(2) It is assumed that in an airgap the flux density is the same as in the iron, and that B is uniform across the airgap (i.e., fringing is neglected).

(3) For each circuit the algebraic sum of the magnetomotive forces across individual branches is equal to the mmf supplied by coils linking the circuit. The magnetomotive force across a branch is the product of H times the length of the branch. A magnetomotive force term is positive in this summation if the reference direction of ϕ is related to the reference direction of the coil current by the right-hand screw rule. (This rule is a statement of Ampere's circuital law.)

(4) At any junction of branches, the algebraic sum of fluxes entering equals the algebraic sum of the fluxes leaving. (This rule is a statement of Gauss' law for magnetic fields.)

(5) In any given branch, H is related to $B = \phi/A$ by the magnetization curve of magnetic material, or in an airgap by

$$H = \frac{B}{\mu_0} = \frac{\phi}{\mu_0 A} \qquad (4\text{-}30)$$

It is noted that rules (3) and (4) are quite similar to the Kirchhoff voltage and current laws.

For examples of solutions, we refer to the three simpler cases shown in Figure 4-25. At (a) there is only one branch, and the circuit has uniform cross section. Therefore, ϕ is everywhere the same, and since H and B are related by a magnetization curve, H is also uniform, and rule (3) gives

$$H\ell = Ni$$

We assume a magnetization curve is available, and from $H = Ni/\ell$ the corresponding B is obtained graphically. From this, $\phi = BA$ is known. Thus, if Ni is given, ϕ can be found. Also, if ϕ is given, B is known, H is determined from the curve, and finally Ni can be found.

The example of Fig. 4-25(b) is more complicated, because there are two branches, and the corresponding equation is

$$H_1\ell_1 + H_2\ell_2 = Ni \qquad (4\text{-}31)$$

Although ϕ is the same in both branches, H_1 and H_2 are different because $B_1 = \phi/A_1$ and $B_2 = \phi/A_2$ are different.

Suppose we are to obtain Ni if a required value of ϕ has been specified. H_1 and H_2 are obtained from the same B-H curve (assuming both branches are of the same magnetic material) as indicated in Figure 4-26(a). B_2 is greater than B_1 because A_1 is greater than A_2. Corresponding values of H_1 and H_2 are read off the curve, and substituted in Eq. (4-31) to give Ni as the required answer.

Now consider the more interesting case, where Ni is given and we are asked to find ϕ. The key to the solution of this problem is in the fact that ϕ must be the same in both branches. Accordingly, as in Figure 4-26(b), a

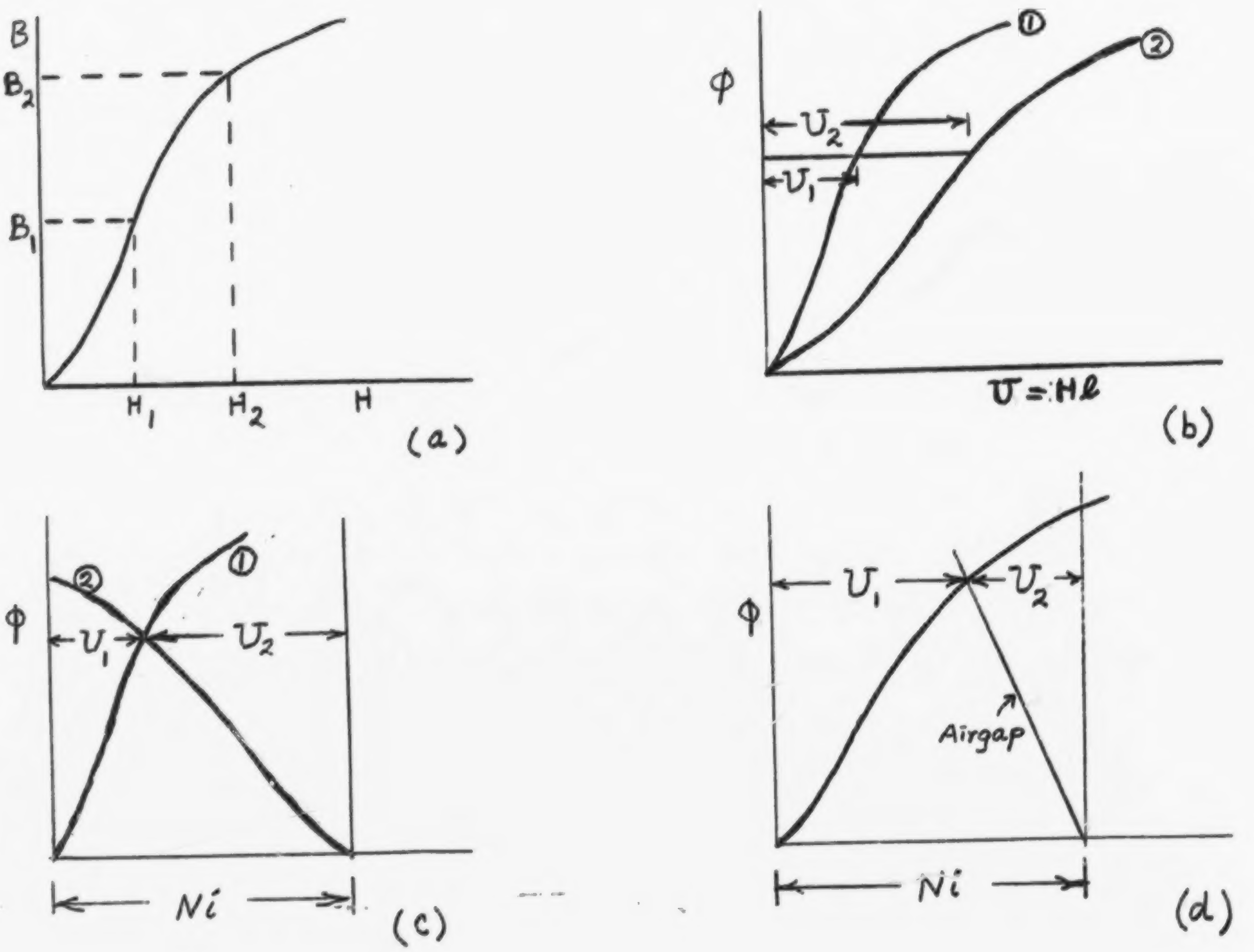


Figure 4-26.

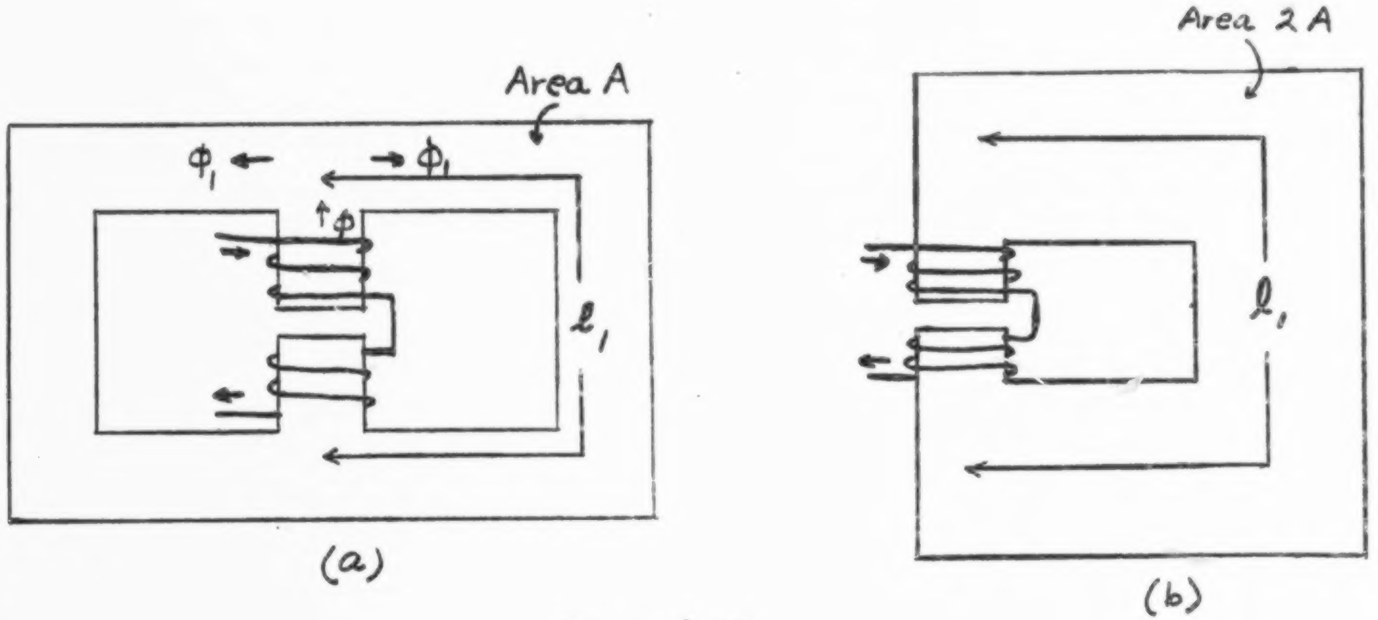
magnetization curve for each branch is constructed as a curve of ϕ vs. magnetomotive force ($H\ell$), and the value of ϕ is found such that the two abscissas add up to the given Ni as required by Eq. (4-31). That is, the horizontal line must be moved up and down until $U_1 + U_2 = Ni$. Note that these two curves are for this specific circuit, since curve (1) was obtained from the universal B-H curve by multiplying B by A_1 and H by ℓ_1, and curve (2) was obtained by using A_2 and ℓ_2 in a similar way. This method involves a graphical procedure, incorporating a series of trials. The trial aspect can be eliminated, however, by using the scheme shown in Figure 4-26(c). One of the curves is reversed, and positioned so that the two origins are spaced an amount Ni. Then, $U_1 + U_2 = Ni$ is automatically satisfied at the point of intersection, which then yields the required value of ϕ.

The above method is restricted to circuits of not more than two branches. In many cases, the second branch is an airgap, as in Figure 4-25(c). This is merely a simple example of the previous case, in which ϕ vs. U for the airgap is a linear relationship

$$U_2 = H_2\ell_2 = \left(\frac{\mu_0 \ell_2}{A}\right)\phi$$

Thus, series circuits of uniform cross section, with an airgap, are easily solved graphically, as in Figure 4-26(d).

Many arrangements with two parallel circuits, like the generator of Figure 4-19(b), can be solved as a single series circuit, because of symmetry. Thus, the two circuits of Figure 4-27 are equivalent as regards the relationship between ϕ and Ni. We can see this by observing that $\phi = 2\phi_1$, in Figure 4-27(a), so that in a side branch $B = \phi/2A$. By increasing the area to 2A in Fig. 4-27(b), B, and therefore H, remain the same.



(a)

(b)

Figure 4-27.

This kind of simplification does not apply, however, to Figure 4-28, because the two sides are not symmetrical. The only way this circuit can be solved is to assume that ϕ_b is known. Then, H_5 can be obtained from properties of the airgap, and magnetization curves yield H_3 and H_4. Thus, in the equation

$$H_2\ell_2 - (H_3\ell_3 + H_5\ell_5 + H_4\ell_4) = 0 \qquad (4\text{-}32)$$

obtained by summing $H\ell$ terms around the outside in accordance with rule (3), the quantity in parenthesis is known. From this H_2 can be determined, and from a magnetization curve B_2 and ϕ_c become known. Then, from rule (4) we have

$$\phi_a = \phi_b + \phi_c \qquad (4\text{-}33)$$

and ϕ_a determines H_1. Finally, the equation

$$H_1\ell_1 + H_2\ell_2 = Ni$$

for the left-hand loop yields the value of Ni. If the problem is stated with the coil mmf as the given quantity, the only way to solve for ϕ_b is to make a series of assumptions of ϕ_b, then to solve for corresponding values of Ni as indicated above. The results can be plotted as a curve of ϕ_b vs. Ni, from which the value of ϕ_b can be read, corresponding to the given Ni.

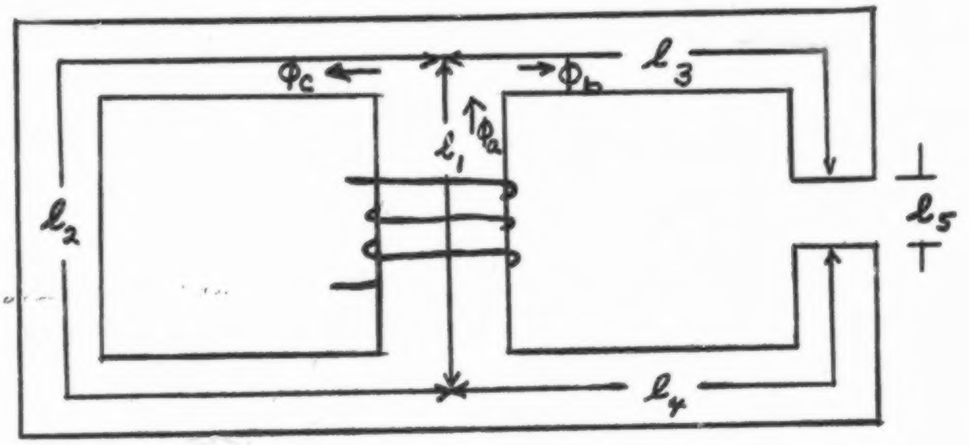


Figure 4-28.

4-11. Self Inductance

In the previous section we have seen how the relationship between flux through a coil and the mmf of the coil can be determined. In general, this relationship is non-linear but when airgaps are present the relationship may be approximately linear for a considerable range of the variables ϕ and Ni. Also, when no iron (or other ferromagnetic material) is present, the relationship is exactly linear, for all practical purposes. For the present, we shall consider that the relationship is linear, and write

$$\phi = P\,Ni \qquad (4\text{-}34)$$

where P is a proportionality factor called the permeance of the magnetic circuit.

Recall now that the emf induced in a coil of N turns linked by a flux ϕ, using the reference directions shown in Figure 4-29, is

$$e_{ab} = -\frac{d(N\phi)}{dt} \qquad (4\text{-}35)$$

This equation has been referred to here to emphasize the importance of the quantity $N\phi$, called the flux linkage of the coil. Using Eq. (4-34) for ϕ, we get

$$N\phi = P\,N^2 i$$

and define a new quantity

$$L = \frac{N\phi}{i} \qquad (4\text{-}36)$$

which, in view of the above expression for $N\phi$, can also be written

$$L = PN^2 \qquad (4\text{-}37)$$

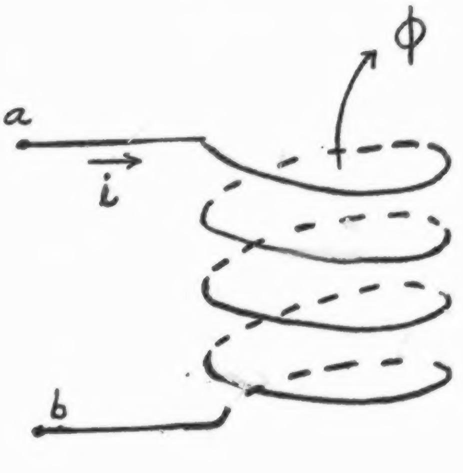


Figure 4-29.

L is a property of the coil, called the self inductance. When current is in amperes and emf is in volts, L is in henrys. The adjective "self" implies that the flux in question is due to current in the coil itself, as distinct from any flux due to external currents or magnets.

As an important observation, note that for a fixed magnetic circuit, Eq. (4-37) shows that inductance is proportional to the square of the number of turns.

Returning now to Eq. (4-35), we may use Li for Nϕ, to give

$$e_{ab} = - \frac{d(Li)}{dt} \qquad (4\text{-}38)$$

as an expression for the emf induced in a coil by its own current.

Of course, Figure 4-29 does not show the magnetic circuit, which can have any degree of complexity, or the coil may be isolated in air. If there is a moving part in the magnetic circuit, or if the dimensions of the coil change with time (i.e., by squeezing together or stretching out the turns as in a coil spring which is being pulled and released) P and therefore L can vary with time. For this reason, L is included under the derivative symbol in Eq. (4-38). In many cases, however, L is constant, and so in that case

$$e_{ab} = - L \frac{di}{dt} \qquad (4\text{-}39)$$

Now we consider the terminal voltage v_{ab} in relation to the current, including the effect of coil resistance R_c. If R_c is negligible, recalling that $v_{ab} = - e_{ab}$, we have

$$v_{ab} = \frac{d(Li)}{dt} \quad \text{or} \quad v_{ab} = L \frac{di}{dt} \qquad (4\text{-}40)$$

The phenomenon of voltage induced in a coil having resistance was treated in connection with Figure 4-14(a). Equation (4-18) applies, except that in that equation (for v_{QP}), current flows from P to Q through the coil. Now we want an equation for v_{ab}, with current flowing through the coil from a to b. Hence, the sign on the i term will change, giving

$$v_{ab} = iR_c + e_{ba} = iR_c - e_{ab}$$

or, inserting Eq. (4-37),

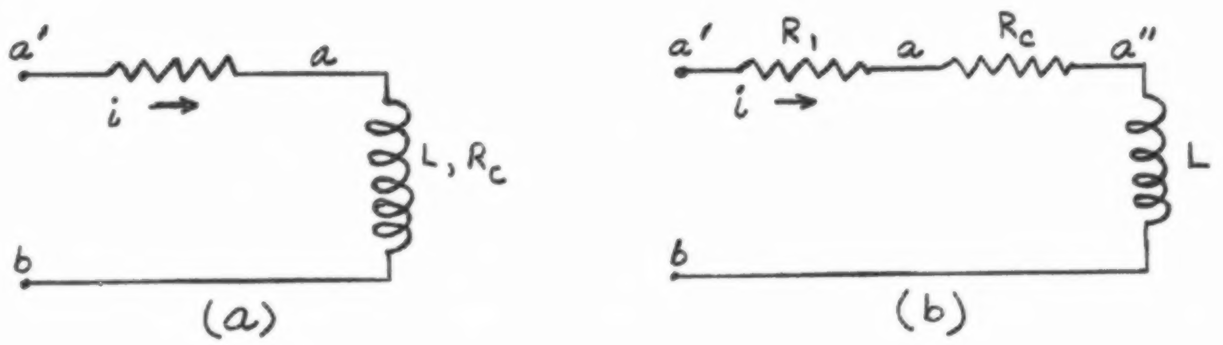
$$v_{ab} = iR_c + \frac{d(Li)}{dt} \qquad (4\text{-}41)$$

The sign on iR_c can be confirmed by imagining the special case where i is constant. Then Eq. (4-40) becomes $v_{ab} = iR_c$, which is consistent with Ohm's law for the reference directions shown.

When a coil is connected to a source through an external resistor, as in Figure 4-30(a), we see that $v_{a'b} = iR_i + v_{ab}$, and so

$$v_{a'b} = i(R_1+R_c) + \frac{d(Li)}{dt} \qquad (4\text{-}42)$$

Thus, the resistance of a coil can be treated like an external resistance, and therefore it is customary to use the equivalent diagram shown in Figure 4-30(b).



(a) (b)

Figure 4-30.

The coil is then regarded as being ideal, having zero resistance. The fictitious voltage

$$v_{a''b} = \frac{d(Li)}{dt}$$

is not accessible to laboratory measurement, but in circuit analysis may be treated as if it were measurable.

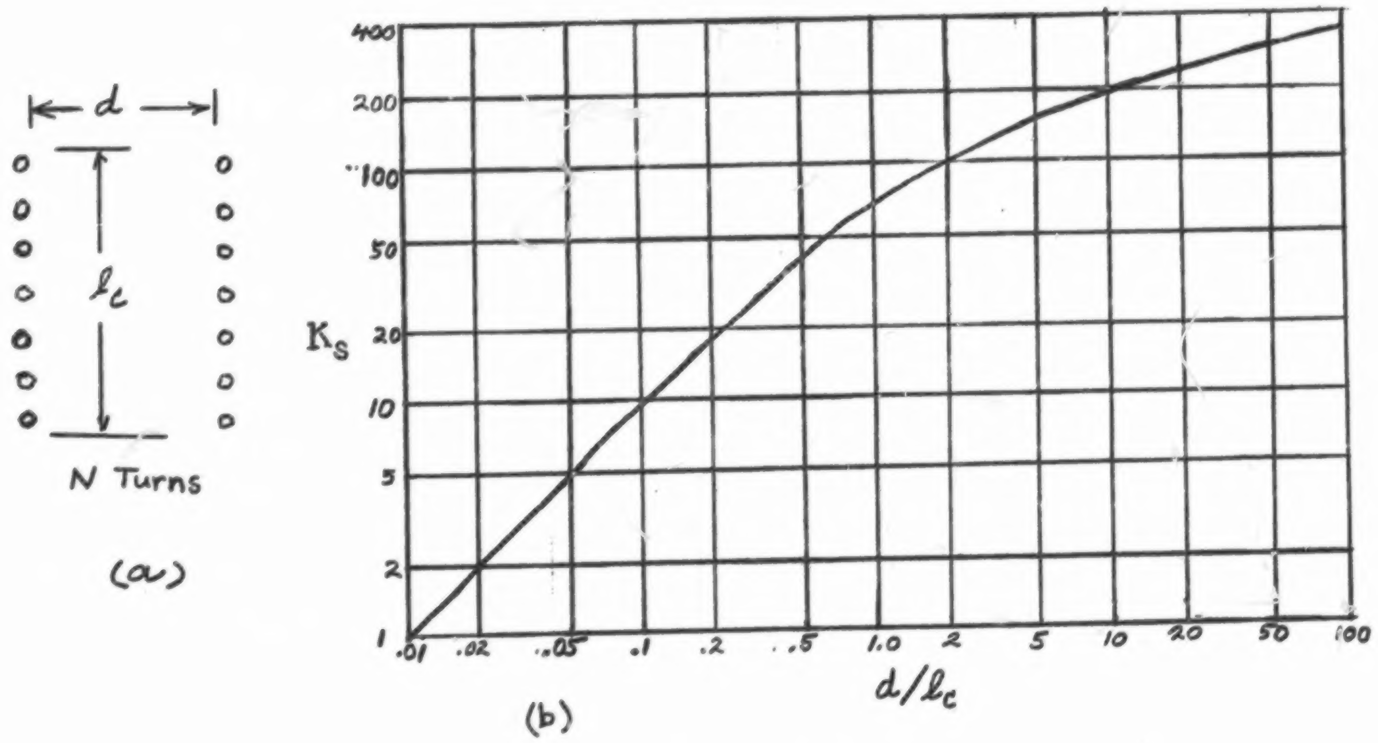
Concerning the computation of inductance, we refer to Eq. (4-36) which shows that if the flux due to a given current can be found by solving the magnetic circuit, that equation will give the inductance. In general, inductance is a function of number of turns, coil dimensions, and sizes and permeabilities of the magnetic circuit components (if there is magnetic circuit). Formulas for coils of various shapes and sizes, in air, will be found in handbooks. Two examples are given below.

For a single layer air coil of length ℓ_c, diameter d, and N turns, as shown in Figure 4-31(a), the inductance is

$$L = K_s d N^2 \qquad \text{(henrys)} \qquad (4\text{-}43)$$

where K_s is the function of the dimeter to length ratio given graphically in Figure 4-31(b). Dimensions ℓ_c and d are in meters. A single circular loop of wire of diameter d, and wire dimeter w, has an inductance given approximately by

$$L = 2\pi d \left(\ln \frac{4d}{w} - \frac{7}{4}\right) \times 10^{-7} \quad \text{(henrys)} \qquad (4\text{-}44)$$



(a)

(b)

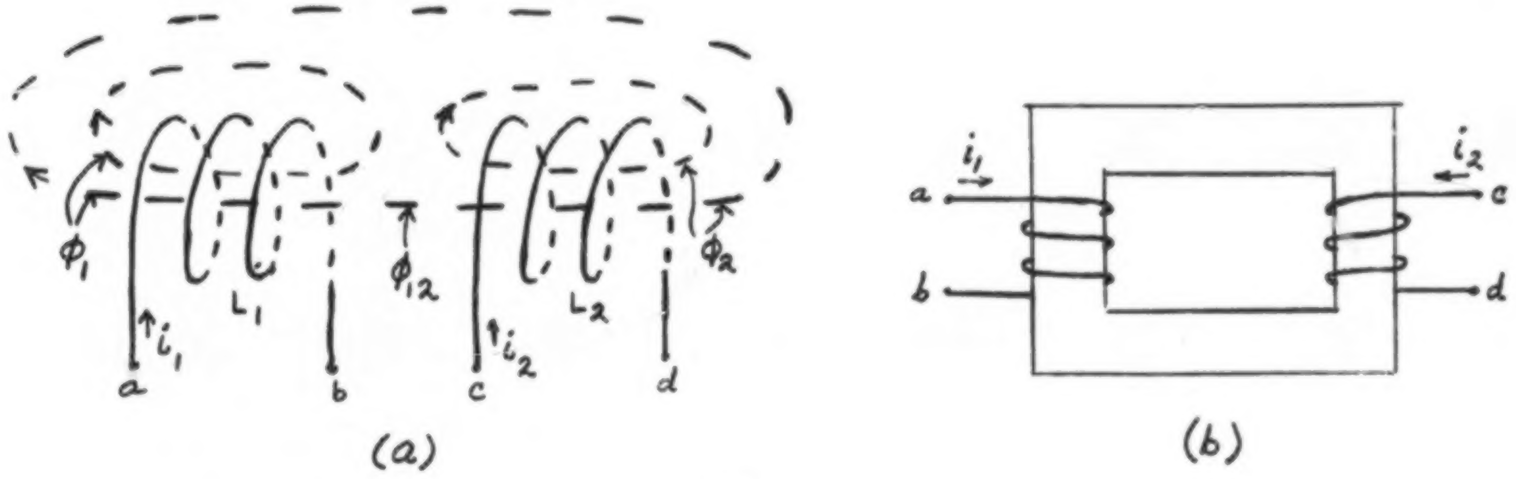
Figure 4-31.*

---

*For a more accurate graph, see Analysis of AC Circuits, p. 399, McGraw Hill Book Co. 1952

4-12. Mutual Inductance

When two coils are in proximity, possible in air, or on a common magnetic core, a changing current in one will cause a changing flux in the other, and will therefore induce a voltage. Examples are shown in Figure 4-32.



(a) (b)

Figure 4-32.

Referring to case (a), let current i_1 be flowing in coil 1, creating flux ϕ_1 through itself and flux

$$\phi_{12} = k_{12}\phi_1$$

through coil 2 (the subscripts 12 imply flux through coil 2 due to current in coil 1). The factor k_{12} is that fraction of flux through coil 1 which also links coil 2, and is always less than unity. An emf will be induced in coil 2, given by

$$e_{cd} = - N_2 \frac{d\phi_{12}}{dt}$$

$$= - N_2 k_{12} \frac{d\phi_1}{dt}$$

However, since $L_1 = N_1\phi_1/i_1$, where L_1 is the inductance of coil 1, we have

$$e_{cd} = - (\frac{N_2}{N_1} k_{12} L_1) \frac{di_1}{dt} \qquad (4\text{-}45)$$

Now suppose the current in coil 1 is zero and that there is a varying current in coil 2. A similar analysis applies, this time involving a fraction k_{21}

1 (that fraction of the flux through coil 2 which also links coil 1). We get

$$e_{ab} = -(\frac{N_1}{N_2}) k_{21}L_2)\frac{di_2}{dt} \qquad (4\text{-}46)$$

2 Experiment shows that the quantities in parentheses in Eqs. (4-45) and (4-46) are the same. This fact is an example of the principle of reciprocity. Thus, both equations can be written

3 $$e_{ab} = -M\frac{di_2}{dt} \quad \text{and} \quad e_{cd} = -M\frac{di_1}{dt} \qquad (4\text{-}47)$$

where M, called the mutual inductance between the coils, stands for either of the quantities in parentheses. M is related to L_1 and L_2 in a manner that can

4 be determined by referring to Eqs. (4-45) and (4-46), to get

$$M = \frac{N_2 k_{12} L_1}{N_1} \quad \text{and} \quad M = \frac{N_1 k_{21} L_2}{N_2}$$

5 Multiplying these two equations together gives

$$M^2 = k_{12}k_{21}L_1L_2$$

6 and finally we have

$$M = k\sqrt{L_1L_2} \qquad (4\text{-}48)$$

7 where $k = \sqrt{k_{12}k_{21}}$ is called the coefficient of coupling. Since k_{12} and k_{21} are each less than unity, k is less than unity, giving the result

$$M \leq \sqrt{L_1L_2}$$

8

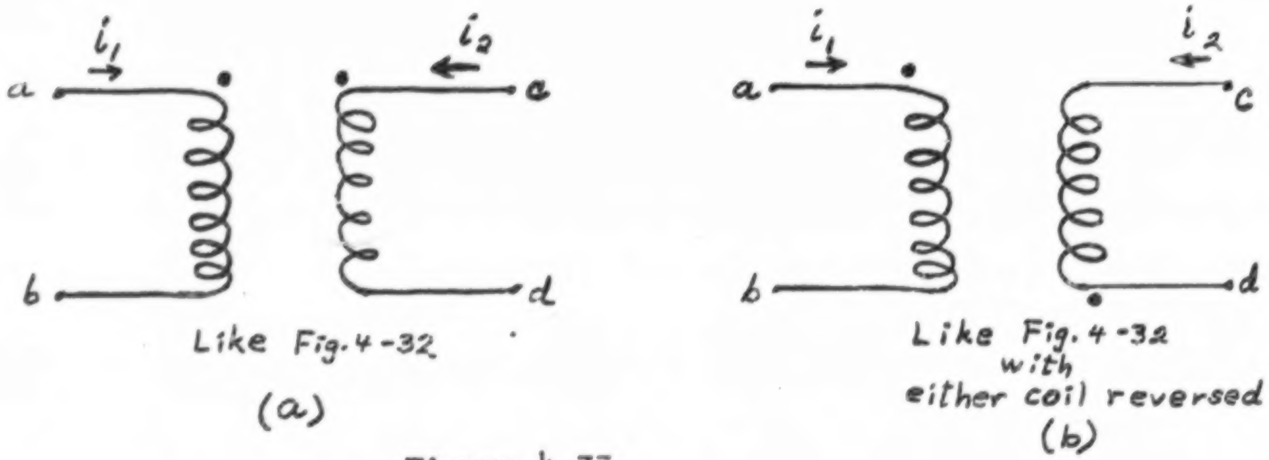
Values of k very close to unity can be attained with iron cores, but in air values of k may be very small. Two coils placed at right angles will have no coupling, and k will then be zero.

9 A device consisting of two or more mutually coupled coils is called a transformer.

The possibility exists that M can vary with time, for example, if the coils are in relative motion. In that case, Eqs. (4-47) become

$$e_{ab} = -\frac{d(Mi_2)}{dt} \quad \text{and} \quad e_{cd} = -\frac{d(Mi_1)}{dt} \qquad (4\text{-}49)$$

Equations (4-47) and (4-49) apply to the arrangements shown in Figure 4-32, where the winding directions are clearly shown. Reversal of the direction of either winding will change all signs to positive. The possibility of a sign ambiguity makes it necessary to show winding directions, or to have a symbol fo signifying what signs to use. In Figure 4-32 each current (i_1 or i_2) enters th terminal (a or c) which appears first in the emf symbol (e_{ab} or e_{cd}). Note tha these reference current directions are magnetically aiding. Each will produce a flux in the same direction through both coils. An alternate system which avoids showing actual winding directions is to place distinguishing marks (usually dots) on a diagram, as shown for two cases in Figure 4-33. These



Like Fig. 4-32
(a)

Like Fig. 4-32
with
either coil reversed
(b)

Figure 4-33.

dots mean that in the actual transformer symbolized by the diagram, currents entering the terminals with dots will be magnetically aiding. The pertinent equations for these two cases are

$$e_{ab} = \mp M \frac{di_1}{dt} \quad \text{and} \quad e_{cd} = \mp M \frac{di_2}{dt} \qquad ;0 < M.)$$

where the - signs are for case (a) and the plus signs for case (b).

Of course, in the above discussion it was assumed that only one coil carr a current at a time. If both coils carry changing currents, self inductance must be included, giving (for the winding directions in Figure 4-32)

$$e_{ab} = - L_1 \frac{di_1}{dt} - M \frac{di_2}{dt}$$

$$e_{cd} = - M \frac{di_1}{dt} - L_2 \frac{di_2}{dt} \qquad (4\text{-}50)$$

Finally, if we include coil resistances, and consider terminal voltages v_{ab} and v_{cd}, we get

$$v_{ab} = R_1 i_1 + L_1 \frac{di_1}{dt} + M \frac{di_2}{dt}$$

$$v_{cd} = M \frac{di_1}{dt} + R_2 i_2 + L_2 \frac{di_2}{dt} \qquad (4\text{-}51)$$

It is not unusual for two mutually coupled coils to be connected in series as in either of the cases shown in Figure 4-34. The total emf ($e_{ad} = e_{ab} + e_{cd}$) will be obtained by adding Eqs. (4-50) and observing that $i_1 = i_2 = i$, giving

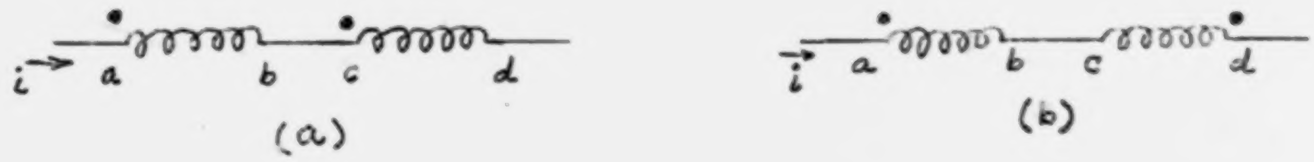


Figure 4-34.

$$e_{ad} = - (L_1 + L_2 + 2M)\frac{di}{dt}$$

For case (b), the signs on both M terms in Eqs. (4-50) will be positive, and thus we then get

$$e_{ad} = - (L_1 + L_2 - 2M)\frac{di}{dt}$$

In each case, the quantity in parentheses is the inductance of the combination, and so we can say that when two coils are connected in series the combined inductance is

$$L = L_1 + L_2 \pm 2M \qquad (4\text{-}52)$$

where the sign depends on the coil connections.

Transformers are among the most important of electrical devices. In power systems they provide means of changing voltage, and also are sometimes used to transmit power to circuits that are conductively isolated. They are also used in communication circuits, in filters and many other applications.

4-13. Transients in R-L Circuits

Consider the circuit shown in Figure 4-35(a). Resistance R includes the coil resistance and any external resistance combined into one value. Assume the switch is closed at time t = 0, at which instant the current in the inductance is zero. As shown in the previous section, the Kirchhoff voltage law for this circuit is

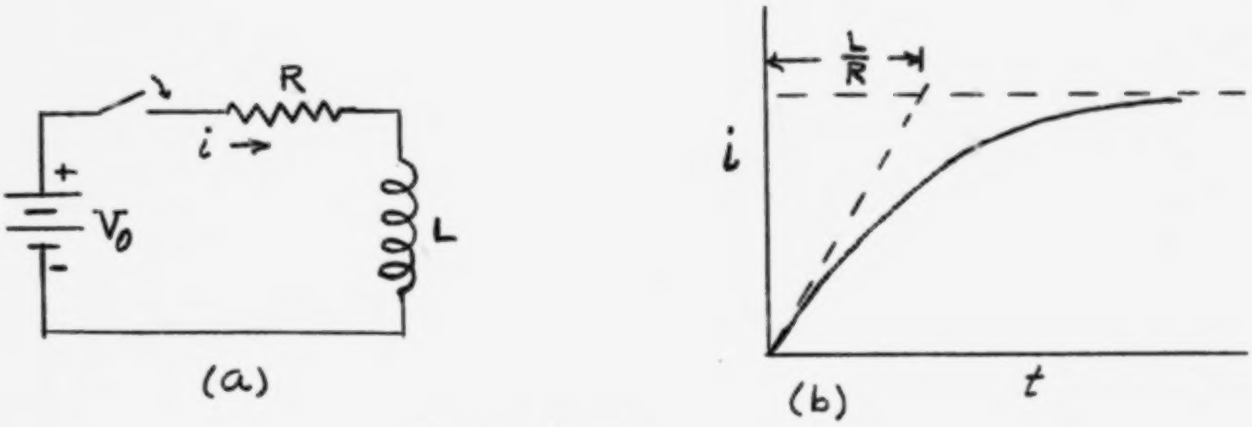


Figure 4-35.

$$V_0 = Ri + L\frac{di}{dt}$$

which can also be written

$$\frac{dt}{di} = \frac{L}{V_0 - Ri}$$

Recognizing the antiderivative, we have

$$t = -\frac{L}{R}\ln(V_0 - Ri) + \ln K$$

$$= -\frac{L}{R}\ln\left(\frac{V_0 - Ri}{K}\right)$$

where K is an arbitrary constant. Multiplying through by -R/L and then writing the inverse of the logarithm to yield an exponential, we get

$$\frac{V_0 - Ri}{K} = e^{-Rt/L}$$

This is readily solved for i, as follows

$$i = \frac{V_0 - Ke^{-Rt/L}}{R} \qquad (4\text{-}53)$$

This is a general solution which includes a constant K. It will be found that the original **Kirchhoff** voltage equation is satisfied by Eq. (4-53) regardless of the value of K. The solution to a specific problem, such as the one stated here, requires a specific value of K, and this is obtained by knowing the value of i at some particular value of t, this knowledge to be obtained from the statement of the problem. With a pair of corresponding values of i and t known, Eq. (4-53) can be solved for K. In most problems i is known at $t = 0$, and this value of i is called an initial condition.

Let us see how the initial condition is determined from a statement of the problem, which included the fact that the current is zero when the switch is closed (that is, for $t < 0$). We recognize that Eq. (4-53) is to apply for $0 \leqq t$, and so we ask whether i as given by Eq. (4-53) should give $i = 0$ when $t = 0$. This would not be true if the current should experience a discontinuous jump as soon as the switch is closed. But such a jump is not possible because of the inductance. We see this by recalling that the voltage across L is $L(di/dt)$, and that with a finite source in the circuit this voltage must remain finite. On the other hand, a sudden jump of current would mean an infinite value of di/dt, and thus it is concluded that a sudden jump is impossible, and that the initial current is zero.

The substitution of $i = 0$ and $t = 0$ in Eq. (4-53) gives

$$K = V_0$$

and so the required solution is

$$i = \frac{V_0}{R}(1 - e^{-Rt/L}) \qquad (4\text{-}54)$$

A graph of this function is shown in Figure 4-35(b). The factor L/R is the time constant, having the geometrical significance indicated. As the exponential term goes to zero, i approaches V_0/R, which is called the steady state value.

The significance of this result is the fact that a finite time interval is required to change the current in an inductor. Ninety per cent of the change will be accomplished when $e^{-Rt/L} = 0.1$, which is approximately when $t = 2.3 \times L/R$. In coils with very large inductance and low R, the time constant can be many seconds.

Perhaps the idea of change is better illustrated by Figure 4-36, where the current initially has a non-zero value V_1/R, attained by virtue of source V_1 having been connected in the remote past. In this case, opening the switch at $t = 0$ inserts an additional source V_2 and resistor R_2. Equation (4-53) applies if V_0 is replaced by $V_1 + V_2$, and R is replaced by $R_1 + R_2$, giving

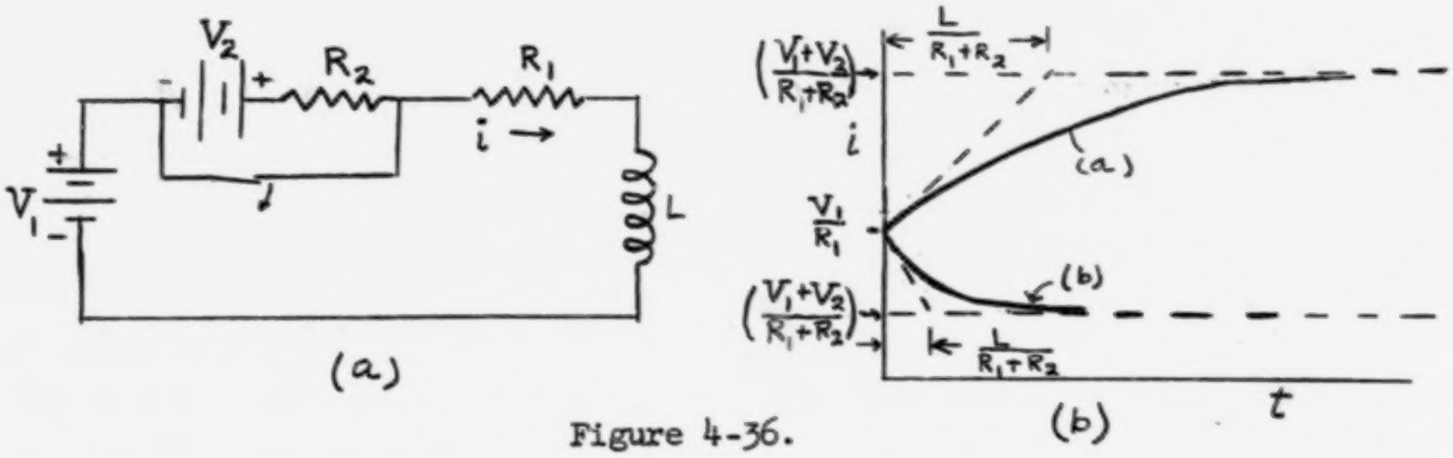


Figure 4-36.

$$i = \frac{V_1 + V_2 - K\,e^{-(R_1+R_2)t/L}}{R_1 + R_2}$$

We again get the initial condition from knowledge that the current cannot change suddenly, and so $i = V_1/R_1$ when $t = 0$, giving

$$\frac{V_1}{R_1} = \frac{V_1 + V_2 - K}{R_1 + R_2}$$

which, when solved for K, gives

$$K = \frac{V_2R_1 - V_1R_2}{R_1}$$

Direct substitution of this value of K yields

$$i = \frac{V_1 + V_2}{R_1 + R_2} - \frac{V_2R_1 - V_1R_2}{R_1(R_1 + R_2)} e^{-(R_1+R_2)t/L}$$

Although **this is a correct** formula, it is not the most convenient form for interpretation because it does not place in direct evidence the initial value and the subsequent change. If we subtract V_1/R_1 from the constant term in the above equation we get

$$\frac{V_1+V_2}{R_1+R_2} - \frac{V_1}{R_1} = \frac{V_2R_1-V_1R_2}{R_1(R_1+R_2)} \quad .$$

which is the same as the coefficient of the exponential in the equation for i. Thus, by adding and subtracting V_1/R_1 the result is

$$i = \frac{V_1}{R_1} + \frac{V_2R_1-V_1R_2}{R_1(R_1+R_2)} \left[1 - e^{-(R_1+R_2)t/L}\right] \qquad (4\text{-}55)$$

By putting the solution in this last form, the initial ($t = 0$) value V_1/R_1 is clearly in evidence, since the bracketed term is zero when $t = 0$. The factor

$$\frac{V_2R_1 - V_1R_2}{R_1(R_1+R_2)}$$

is the total change in current, and can be either positive or negative depending on the relative magnitudes of V_2R_1 and V_1R_2. Both cases are illustrated in Figure 4-36(b). Similarity with Figure 4-35 is evident if it is noted that in each case a factor of the form $(1-e^{-\alpha t})$ multiplies a factor which is the total change of the current. This principle can be used to solve a wide variety of switching problems in which there is a single inductor. Initial and final values of currents can be found from simple d-c circuit principles. The value of α is obtained by reducing all sources to zero, finding the equivalent resistance in series with the inductance, and dividing this by L.

An important special case is shown in Figure 4-37. Here, as a result of connection of the battery in the remote past, a current $V_0/(R_1+R_2)$ is flowing in the inductor. At $t = 0$ the switch is closed, reducing the voltage between A and B to zero. Subsequent to $t = 0$, the current i changes from $V_0/(R_1+R_2)$ to zero. Thus, the equation for the current is

$$i = \frac{V_0}{R_1+R_2} - \frac{V_0}{R_1+R_2}\left(1 - e^{-R_2 t/L}\right)$$

Note that a negative sign appears on the second term because the change is a decrease, and also note that after the switch is closed, only R_2 is in series with L. The above reduces to the simpler form

$$i = \frac{V_0}{R_1+R_2} e^{-R_2 t/L} \qquad (4\text{-}56)$$

This can also be obtained directly from Eq. (4-53) by obtaining the appropriate value for K.

These sample solutions have all been given to find current. Of course, once i is known, the voltage across the inductor can readily be obtained from L di/dt.

4-14. Flux Linkage Theorem

In the previous section it was pointed out that L di/dt must remain finite, because voltage across an inductor remains finite, and therefore that the current in an inductive circuit cannot experience a sudden jump. However, this principle applies only if L is constant with time, and if there is no changing flux due to external causes.

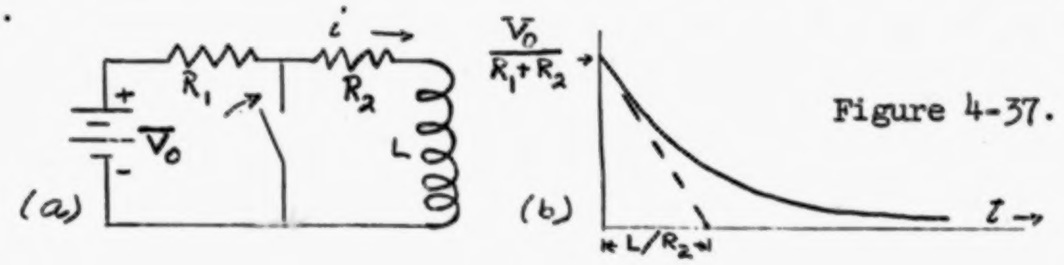


Figure 4-37.

Figure 4-38 shows the more general case which we shall consider briefly.

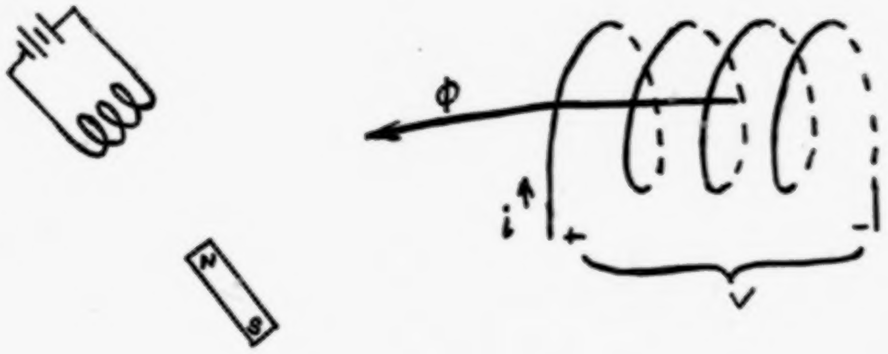


Figure 4-38.

The flux ϕ through the coil is due to current i, and also is in part due to external magnets or current-carrying coils. Thus, the total flux can be written

$$\phi = \phi_e + \frac{Li}{N} \qquad (4\text{-}57)$$

where ϕ_e is due to causes external to the coil, and Li/N is the flux due to i, as determined from Eq. (4-42). The voltage v is

$$\begin{aligned} v &= \frac{d(N\phi)}{dt} \\ &= \frac{d}{dt}(N\phi_e + Li) \end{aligned} \qquad (4\text{-}58)$$

In this case, again using the fact that v must remain finite, it is concluded that the quantity in parentheses cannot experience a sudden jump. However, any one of the quantities within the parentheses can change suddenly, provided there is a compensating change in one of the others.

Two cases are of practical importance. First, suppose there is a sudden change in the magnetic circuit associated with the coil, causing a sudden change ΔL in the inductance, and assume ϕ_e remains constant. Since $N\phi_e + Li$ must remain continuous (be free of a jump) it follows that there must be a corresponding sudden change in i such that

$$(L + \Delta L)(i + \Delta i) = Li$$

and consequently

$$\Delta i = -\left(\frac{\Delta L}{L+\Delta L}\right) i \qquad (4\text{-}59)$$

As a second example, suppose ϕ_e suddenly changes by an amount $\Delta\phi_e$ due to a sudden mechanical displacement of one of the external flux sources. Then we have

$$N\phi_e + Li = N(\phi_e + \Delta\phi_e) + L(i + \Delta i)$$

and

$$\Delta i = -\frac{N}{L}\Delta\phi_e \qquad (4\text{-}60)$$

The principle discussed here, that the total flux linking a coil $(N\phi_e + Li)$ cannot have sudden jumps is known as the constant flux linkage theorem.*

This theorem is useful in solving circuit problems in which certain types of switching operations cause sudden changes in inductance. Consider, for example, Figure 4-39(a) in which there are two inductances which are not magnetically coupled. Assuming the current has reached the steady state value V/R, at time t = 0 the switch across one inductance is closed, suddenly reducing the total inductance from 3L to 2L. In other words, $\Delta L = -L$, and so from Eq. (4-59) we find that the current increases an amount

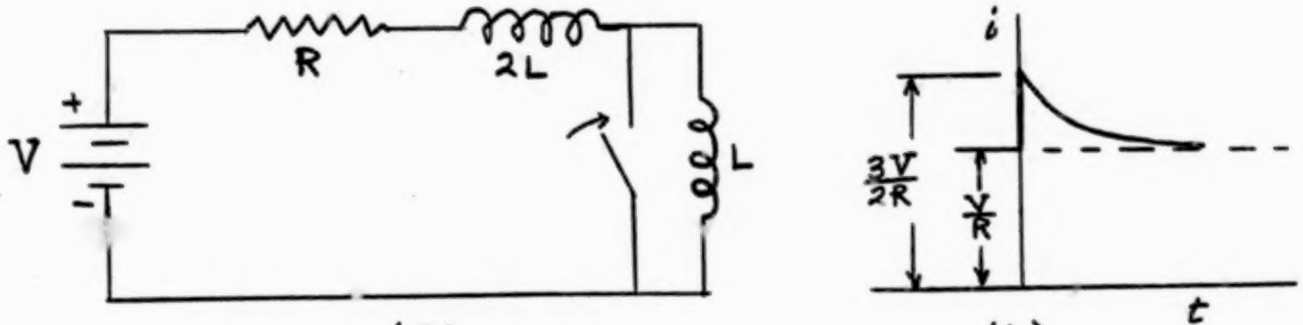
$$\Delta i = -\left(\frac{-L}{3L-L}\right)\frac{V}{R} = \frac{V}{2R}$$

Adding this change to the original current gives (3/2)(V/R) as the current immediately after the switch is closed.

Subsequent behavior of this circuit can be analyzed by methods given in the previous section, using (3/2)(V/R) as the initial value. From inspection of the circuit it is evident that the current will gradually go back to V/R along an exponential curve of time constant L/2R, as shown in Figure 4-39(b).

---

*In this statement the word "constant" means constant over a short period of tim Of course, flux linkage can change, but it must do so smoothly, in a continuous fashion, to use mathematical terminology. It would be better to call this the continuous flux linkage theorem.



(a) (b)

Figure 4-39.

4-15. Energy Stored in a Magnetic Field

Consider a torroidal inductor with iron core, as in Figure 4-40, and assume the core diameter d is small enough to allow the assumption of uniform B over the cross section, and assume the coil resistance is negligible. The power delivered to the coil during a change of current is

$$v_{ab}\, i = (N \frac{d\phi}{dt})\, i$$

Now suppose that in time interval from t_1 to t_2 the flux changes from ϕ_1 to ϕ_2. The energy delivered from the source will be the time integral of power, namely

$$W = \int_{t_1}^{t_2} v_{ab} i dt = N \int_{t_1}^{t_2} i \frac{d\phi}{dt} dt = \int_{\phi_1}^{\phi_2} Ni\, d\phi$$

Now observe that, in terms of B and H, we have

$$\phi = BA \qquad \text{and} \qquad Ni = H\ell$$

so that we get

$$W = \ell A \int_{B_1}^{B_2} H\, dB \qquad (4\text{-}61)$$

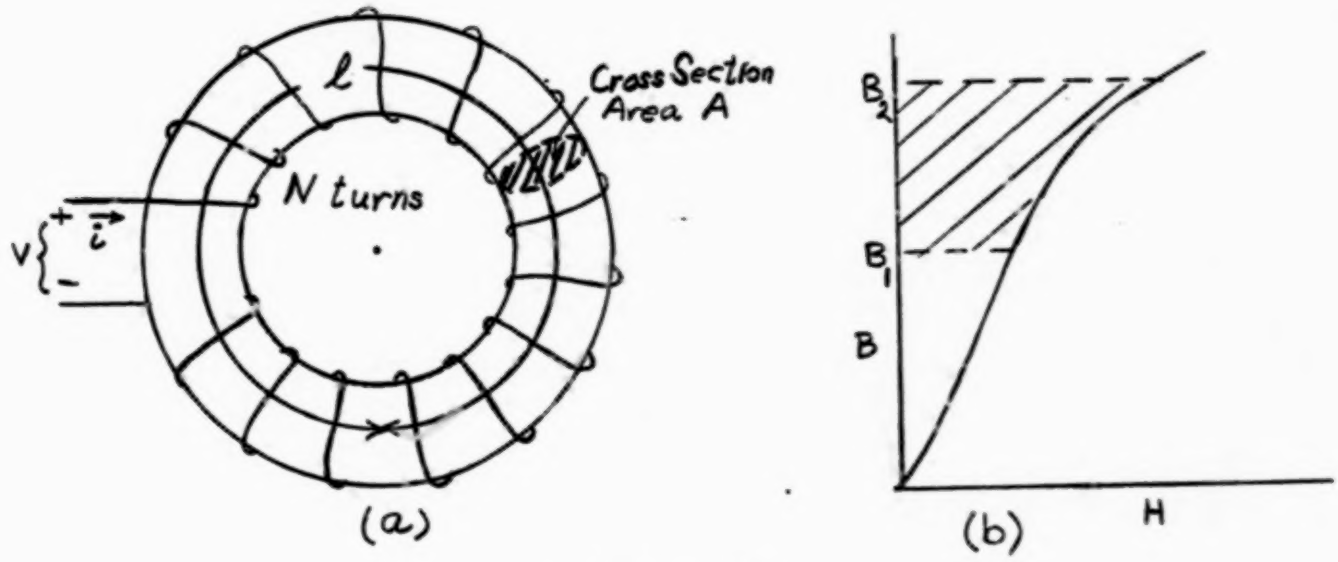


Figure 4-40.

This integral is interpreted as an area in Figure 4-40(b), for the case where the magnetism has been increased from B_1 to B_2. If $B_1 = 0$, we get

$$w = \int_0^B H \, dB \qquad (4\text{-}62)$$

which is interpreted as the energy density in the magnetic field when its strength is B. Actual energy is then obtained by multiplying the density by the volume ℓA. In accordance with this result for the special case of a torroid it is postulated that Eq. (4-62) gives the energy density for all field configurations.

No assumption as to linearity was made in arriving at Eqs. (4-61) and (4-62). Suppose magnetism is carried partially around a hysteresis loop, from O to P to Q, in Figure 4-41. In going from O to P the energy density increases

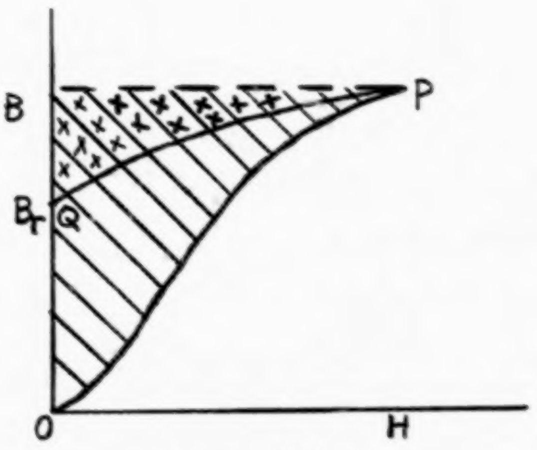


Figure 4-41.

by an amount

$$\int_0^{B_1} H\,dB$$

which is represented by the shaded area. In reducing B to B_r, the energy density will change by

$$\int_{B_1}^{B_r} H\,dB = -\int_{B_r}^{B_1} H\,dB$$

which is a decrease equal to the area covered by the small crosses. This energy is returned to the circuit. Thus, an amount of energy represented by the area between the curves has not been returned, even though the current has returned to zero. This represents an energy loss due to the molecular interactions which cause hysteresis. When magnetism is carried completely around a hysteresis loop, energy is lost in proportion to the area of the loop. This energy loss is an important consideration of the design of generators, motors, and transformers, since in all of these there are repeated reversals of flux.

When the magnetic material is linear, so that we may write $B = k_m\mu_0 H$, Eq. (4-62) reduces to

$$w = \frac{1}{k_m\mu_0}\int_0^B B\,dB = \frac{B^2}{2k_m\mu_0} = \frac{BH}{2} \qquad (4\text{-}63)$$

Energy storage in linear media can also be expressed in terms of inductance. Using the reference directions of Figure 4-42, if the current increases from 0 to i,

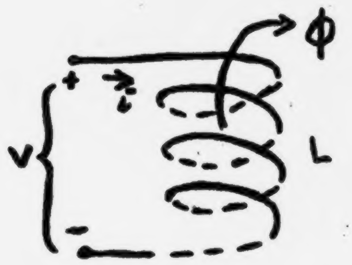


Figure 4-42.

as to goes from 0 to t, for the energy we have

$$W = \int_0^t i\,L\,\frac{di}{dt}\,dt = L\int_0^i i\,di$$

or

$$W = \frac{1}{2} L\,i^2 \qquad (4\text{-}64)$$

We can get the same result from Eq. (4-63), for the torroidal case. The total energy is

$$W = \frac{BA\ H\ell}{2} = \frac{\phi N i}{2}$$

but $\phi N = Li$, from Eq. (4-42), and so Eq. (4-64) is obtained from the above.

4-16. Force Across an Airgap

Airgaps in magnetic circuits experience forces which tend to cause them to close. This force is utilized in a variety of applications, for example in electric relays.

The idea of energy density can be used to determine the force tending to close an airgap. Referring to Figure 4-43, suppose the gap is small, so that there will be negligible error in assuming B lines are parallel in the gap, and that B is negligible outside the gap. The pole face has an area A, and the gap length is ℓ, giving

$$W = \frac{A\ell B^2}{2\mu_0} \qquad (4\text{-}65)$$

for the energy stored in the gap. Now suppose the gap is increased an amount $\Delta\ell$ by the application of an external force F. To do this in practice, there must be a hinge in the iron circuit; otherwise, the force will also have to overcome stress in the iron. Thus, F is only that force required to move the pole face against magnetic force.

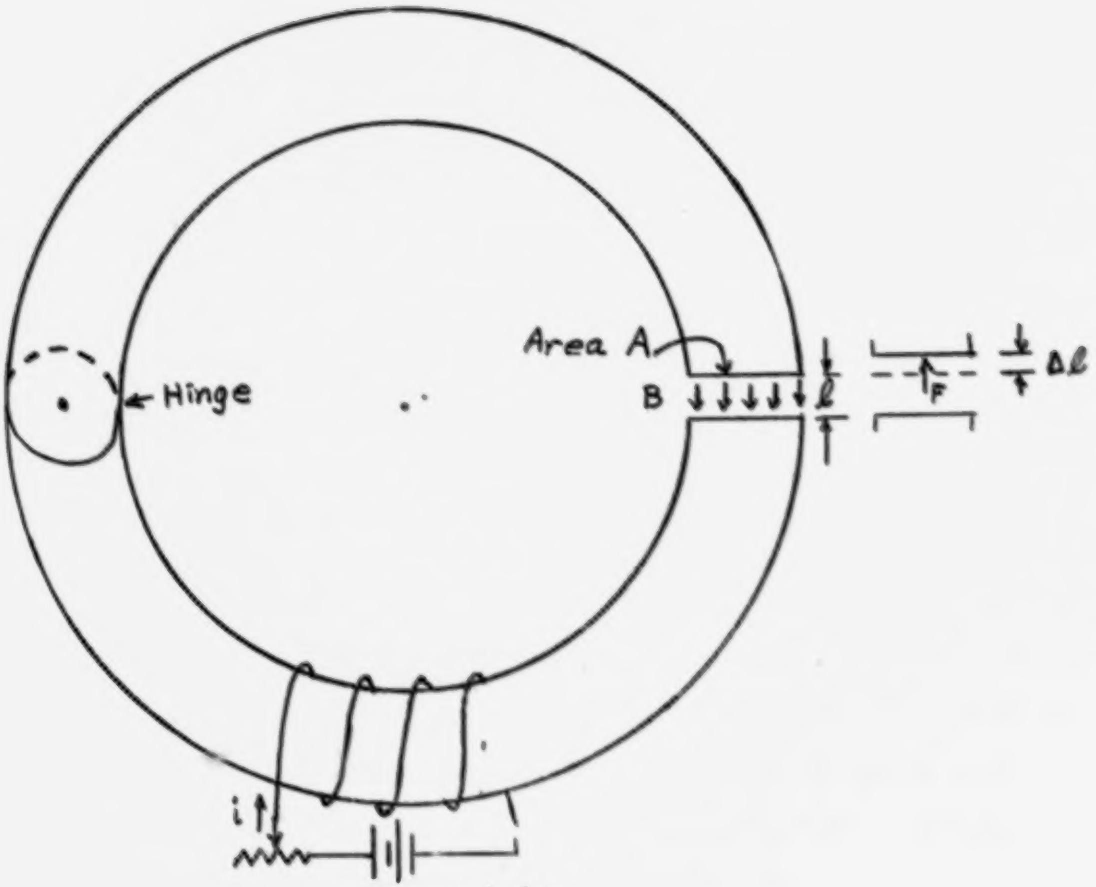


Figure 4-43.

If conditions of the external circuit remain constant, increasing the airgap would cause a reduction of ϕ which would complicate the analysis. To avoid this, assume that,coincident with the change in ℓ, i is gradually increased by adjustment of Resistor R, so as to maintain ϕ constant. All mechanical energy expended will then to into the energy stored in the field.* This change is

$$\Delta W = \frac{AB^2}{2\mu_0} \Delta\ell$$

But from mechanics we know that this energy is also $F\Delta\ell$, and so we have

$$F = \frac{AB^2}{2\mu_0} \qquad (4\text{-}61)$$

as the expression for the required force.

---

*No energy interchange occurs with the battery due to change in ℓ because $d\phi/dt = 0$, and so the coil voltage is zero.

Chapter 5

STEADY STATE CIRCUIT ANALYSIS

Introduction

We begin this chapter with a description of the experiment illustrated in Fig. 5-1a, in which an alternating current generator is suddenly connected to a circuit consisting of R and L. Oscilloscopic records are taken of the voltage and current, yielding curves as shown in Fig. 5-1b. The voltage wave begins at an arbitrary point, t = 0, depending on the instant the switch was closed. The form of the current wave will depend upon this instant, and so the experiment will not yield the same result with each trial. Another example is shown in Fig. 5-1c. However, after a period of time, the current wave settles down to a form called the steady state wave. If the steady state wave is projected back to t = 0, as indicated by the dashed curves, it is found that the actual current can be viewed as the sum of this wave and the curve labelled transient. Whenever

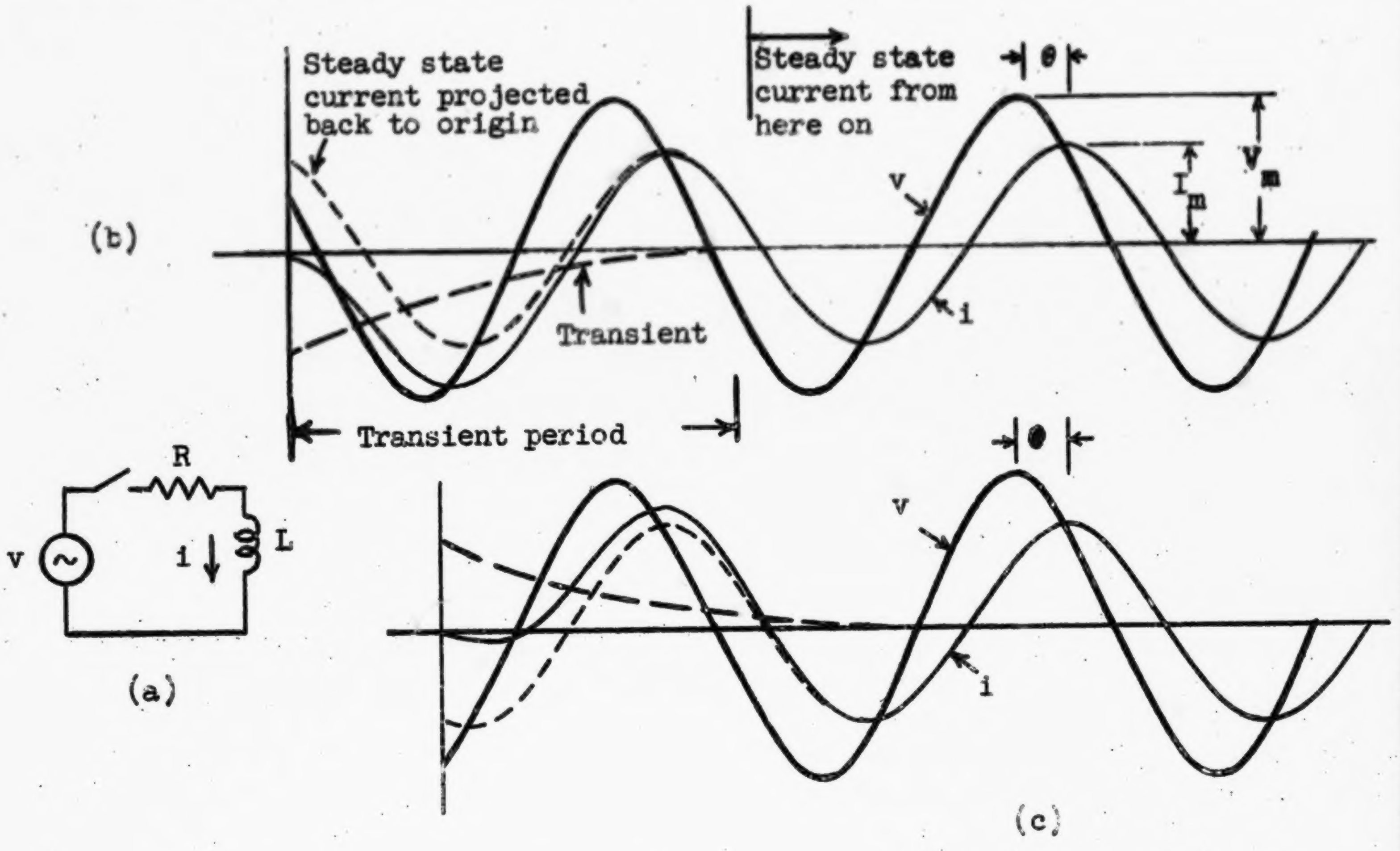

Steady state current projected back to origin
Steady state current from here on
θ
V_m
I_m
v
i
(b)
Transient
Transient period
R
L
v
i
(a)
θ
v
i
(c)

Figure 5-1.

the network includes resistance elements, as in the case under consideration, the transient will die out. In more complicated networks the form of the transient may be more complicated than the simple exponential shown.

This brief analysis illustrates the following:

(1) When a sinusoidal source is suddenly applied to a circuit, there is a transient period during which the response is not a repeating wave. (In the example above, the response was the current.)

(2) The actual form of the response during the transient period depends on the instant at which the switch is closed.

(3) After the transient has died out, the wave approaches a steady state wave having a magnitude and position on the time axis, relative to the source wave, which are independent of the time of switching. Thus, referring to (b) and (c) of Fig. 5-1, the ratio V_m/I_m and the phase difference, θ, are the same for both.

Although this has been presented as a description of an imagined experiment, theoretical confirmation of the conclusions is possible. However, such theoretical confirmation must be deferred until circuit analysis has been studied quite thoroughly, and therefore will not be considered at this time.

The fact that the steady state response is independent of what takes place during the transient period makes steady state analysis possible.

One may legitimately question whether a study of the steady state response of a network has practical value. In replying to this question, either of two different answers can be given, depending on whether a network is designed primarily to transmit power or communication signals. Power transmission usually takes place at a nearly constant rate, and so steady state analysis provides most of the needed answers. On the other hand, a communication system can be thought of as a power system in which signals are coded in the form of changes in the power transmitted (i.e., the turning on of a light can be a signal). Thus, transient considerations must also enter into the analysis of communication circuits. However, the analysis of networks for steady state response provides an important foundation for the study of both power and communication systems.

5-1. Description of a Sinusoid

Consider a plot of the function

$$i = I_m \cos(\omega t + \alpha) \qquad (5\text{-}1)$$

which is shown in Fig. 5-2a. It is labelled as a current, but a similar expression could be written for a voltage or other quantity. In this equation, I_m, ω, and α are constants, and t is the variable time. Since ω is constant, ωt can also be regarded as a variable. Both these variables (t and ωt) are plotted along the horizontal axis in Fig. 5-2a. Since time has no beginning, it is necessary to choose some arbitrary instant as $t = 0$. At this instant, Eq. (5-1) gives $i = I_m \cos \alpha$, as indicated. Since α determines the value of $(\omega t + \alpha)$ when $t = 0$, it is called the initial angle of the wave. I_m is the maximum value, or amplitude.

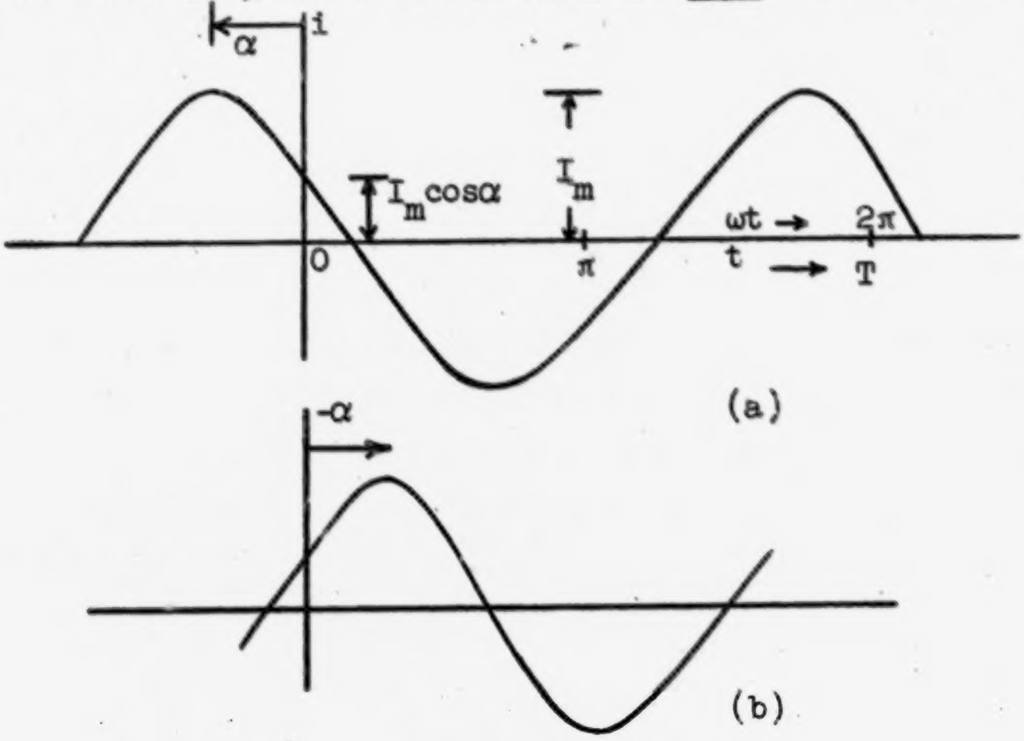


Figure 5-2.

To interpret the parameter ω, note that, since the cosine of an angle is unchanged if 2π is added to the angle, the wave will repeat itself when ωt has increased by 2π. Thus, the wave is said to be periodic. Let T be the increase in time t required to make ωt increase by 2π, so that

$$\omega T = 2\pi$$

or

$$\omega = \frac{2\pi}{T}$$

T is the period of the wave; and $1/T = f$, called the frequency, is the number of times the wave repeats in one second. (Earlier, the symbol T was used to denote the time constant of an RC or RL network. Now, the same symbol is being used for the period of a periodic function.) For example, if $T = 0.1$ sec., there will be 10 repetitions of the wave in one second. In terms of f, Eq. (5-2) becomes

$$\omega = 2\pi f \tag{5-3}$$

Thus, ω has been established as a parameter related to the period (or frequency) of the wave. In view of ωt being an angle, ω is called angular frequency.

If I_m, α and ω are known as numerical values, the corresponding wave can be plotted unambiguously. Thus, these three parameters completely specify a wave. Furthermore, ω is determined by the frequency of the source; if this is given, the solution of steady state problems amounts to the finding of values of magnitudes and initial angles of various voltage and current waves in a network.

The decision to write Eq. (5-1) as a cosine is entirely arbitrary. If we let $\alpha = \alpha' - \frac{\pi}{2}$, Eq. (5-1) becomes

$$I_m \cos(\omega t+\alpha' - \tfrac{\pi}{2}) = I_m \sin(\omega t+\alpha')$$

Thus, any such wave can be written as a cosine function or a sine function. The cosine form has certain advantages, and so we shall use it in this text. Regardless of whether it is written as a cosine or sine function, the wave we are considering is called a sinusoid, and its shape is said to be sinusoidal.

Although the magnitude parameter I_m is entirely adequate to determine the size of a sinusoid, another value (called the rms value) is usually employed. This value, which we shall identify as I_{rms} or simply I, is the square root of the average (mean) value of the square of the wave. (The symbol rms is an abbreviation for the words "root", "mean", "square".) For simplicity, consider the current wave (with zero initial angle)

$$i = I_m \cos \omega t$$

for which the squared wave is

$$i^2 = I_m^2 \cos^2 \omega t$$

$$= \frac{I_m^2}{2}(1 + \cos 2\omega t)$$

The second term shows that the i^2 curve varies in such a way as to have equal areas above and below the dashed line in Fig. 5-3. Therefore, this dashed line is the average value of the i^2 curve,* and so we can write

$$\text{Average of } i^2 = \frac{I_m^2}{2}$$

and the square root of this, the rms value, is

$$I = \frac{I_m}{\sqrt{2}} \qquad (5\text{-}4)^{**}$$

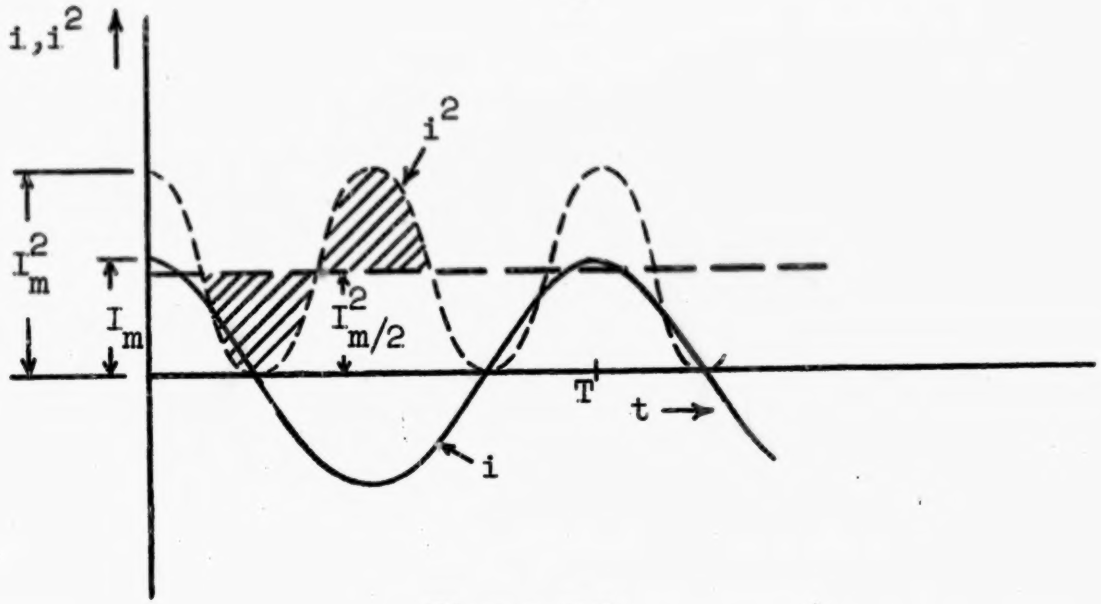


Figure 5-3.

For an interpretation of the rms value, assume $i = I_m \cos \omega t$ is the current in a resistor R. At any instant of time, power will be dissipated at a rate i^2R, and from the above proof that the average of i^2 is $I_m^2/2$, it follows that the average power dissipation will be $(I_m^2/2)R$. In view of Eq. (5-4) this is also I^2R. Thus, for finding average power, I can be used in the I^2R formula exactly as if it were a d-c current. Since I has this interpretation in terms of the average effectiveness of the wave in producing heat in a resistor, I is sometimes called the *effective* value of the wave. The rms designation is to be preferred, however.

---

*Strictly speaking, the areas above and below the dashed curve are equal only if reckoned over an integral multiple of the period of the wave. For example, the dashed line would not be the average reckoned over a period and a half, but it can be shown that over a long time interval (many periods) the average is only slightly affected by the inclusion of a fraction of a period. That is, the averages over 100 periods and over 100 + 1/2 periods are very nearly the same, as can be seen by making a simple calculation.

**Rms values can be defined for waveshapes which are not sinusoidal. However, the $1/\sqrt{2}$ relationship does not then generally apply. Equation (5-4) is given here for a sinusoidal wave.

Most a-c instruments for measuring voltage and current are calibrated to indicate rms values. Since the rms value is used more prevalently than the amplitude, in this chapter we shall use the rms value in writing an equation for a sinusoid. Accordingly, we write Eq. (5-1) in the alternate form

$$i = \sqrt{2}\, I \cos(\omega t + \alpha) \qquad (5\text{-}5)$$

5-2 Complex Numbers

Complex numbers were invented by mathematicians to generalize the theory of quadratic equations. For example, the equation

$$x^2 - 2b\,x + c = 0$$

has roots given by

$$x_1 = b + \sqrt{b^2 - c} \quad , \; x_2 = b - \sqrt{b^2 - c}$$

If $b^2 \geq c$, x_1 and x_2 are real numbers. If $b^2 < c$, the square root of a negative number results. But there is no *real* number whose square is negative. Nevertheless, these same formulas for the roots can still apply if an *imaginary unit* j is *defined* such that

$$j^2 = -1 \qquad (5\text{-}6)^*$$

Then, if $b_2 < c$, the two roots can be written

$$x_1 = b + j\sqrt{c - b^2} \quad , \qquad x_2 = b - j\sqrt{c - b^2}$$

Each of these is a *complex number* consisting of a *real part* b and an *imaginary part* $\sqrt{c-b^2}$. But it is one thing merely to write $b \pm j\sqrt{c-b^2}$, and another to know what it means. In mathematics an "absolute" meaning is not required; meaning is

---

[*] The symbol i is used in mathematics, but due to the extensive electrical engineering literature in which i is used for current, the symbol j is standard in electrical engineering.

embodied in the rules of operations which are defined for a quantity. The rules of operating on a complex number are defined in such a way as to make them useful. For example, in order for x_1 to satisfy the given equation, it is necessary to define what is meant by x_1^2 and $-2bx_1$, when $x_1 = b + j\sqrt{c-b^2}$. The required definition is that a complex number shall be treated like an ordinary binomial, with j^2 being replaced by -1 whenever j^2 occurs. Thus

$$x_1^2 = (b + j\sqrt{c-b^2})^2 = b^2 + j^2(c-b^2) + j2b\sqrt{c-b^2}$$

$$= 2b^2 - c + j2b\sqrt{c-b^2}$$

and

$$-2bx_1 = -2b^2 - j\,2b\sqrt{c-b^2}$$

The sum of these is -c, and so it is seen that $x^2-2bx + c = 0$ is satisfied for $x = x_1$.

Let the general notation

$$\bar{A} = A_1 + jA_2$$

$$\bar{B} = B_1 + jB_2 \qquad (5\text{-}7)^*$$

represent two complex numbers. (In this text a bar above a symbol will be used to denote that it is complex.) For reference, we shall state the rules of algebraic operations in terms of these examples. But first we must define what is meant by equality. By $\bar{A} = \bar{B}$ we mean that the real parts are equal to each other and the imaginary parts are equal to each other. That is,

$$A_1 = B_1 \qquad \text{and} \qquad A_2 = B_2$$

---

[*] A complex number is merely a combination of two numbers, which cannot themselves be combined. A good way to regard this pair of numbers is

$$\bar{A} = (1)A_1 + (j)A_2$$

This shows a symmetry between the two parts. One part is $A_1 \times$ (the real unit 1), and the other part is $A_2 \times$ (the imaginary unit j). A_1 and A_2 are real.

Conversely, if $A_1 = B_1$ and $A_2 = B_2$, we can write $\bar{A} = \bar{B}$. In other words, two complex numbers are equal if and only if their real parts are equal and their imaginary parts are equal.

For the operations of addition, multiplication and division, using the ordinary rules for binomials, we get

$$(A_1 + jA_2) + (B_1 + jB_2) = (A_1 + B_1) + j(A_2 + B_2)$$

$$(A_1 + jA_2)(B_1 + jB_2) = A_1B_1 - A_2B_2 + j(A_2B_1 + A_1B_2) \qquad (5\text{-}8)$$

$$\frac{A_1 + jA_2}{B_1 + jB_2} = \frac{A_1B_1 + A_2B_2 + j(A_2B_1 - A_1B_2)}{B_1^2 + B_2^2}$$

The validity of the formula for division can be confirmed by showing that an identity results when both sides of the equation are multiplied by $B_1 + jB_2$. The formula for division is rather complicated, but it can be reconstructed by multiplying numerator and denominator by the number $B_1 - jB_2$, as follows:

$$\frac{\bar{A}}{\bar{B}} = \frac{A_1+jA_2}{B_1+jB_2} = \frac{(A_1+jA_2)(B_1-jB_2)}{(B_1+jB_2)(B_1-jB_2)} = \frac{A_1B_1 + A_2B_2 + j(A_2B_1-A_1B_2)}{B_1^2 + B_2^2}$$

The number $B_1 - jB_2$ used here is related to $\bar{B}$ by having the sign of the imaginary part changed. This number is called the complex conjugate of $\bar{B}$, and will be designated by $\bar{B}^*$. Using the conjugate, it follows that:

$$\bar{B}\,\bar{B}^* = B_1^2 + B_2^2$$

$$\frac{\bar{B} + \bar{B}^*}{2} = B_1 \qquad (5\text{-}9)$$

$$\frac{\bar{B} - \bar{B}^*}{2j} = B_2$$

The first of these follows directly from the rule for multiplication; the second and third are readily obtained by substituting $\bar{B} = B_1 + jB_2$ and $\bar{B}_1^* = B_1 - jB_2$. Thus, by combining a complex number with its conjugate, the real or imaginary part, or the sum of their squares can be obtained. (Each of these is a real number.)

The relationship

$$e^{j\theta} = \cos\theta + j\sin\theta \qquad (5\text{-}10)^{*}$$

is called *Euler's* identity. The right-hand side of this equation is the *definitio* of the symbol $e^{j\theta}$. Thus, the formula requires no proof, but properties of $e^{j\theta}$ so defined must be investigated. The properties of $e^{j\theta}$ we shall need are briefly treated in the footnote, where it is shown that $e^{j\theta}$ can be multiplied and differentiated like a real exponential.

Euler's identity can be used to provide an alternate notation for a complex number

$$\bar{A} = A_1 + jA_2$$

This can be shown by multiplying both sides of Eq. (5-10) by a *real number* A, to give

$$Ae^{j\theta} = A\cos\theta + jA\sin\theta$$

In view of the meaning of equality of complex numbers, we see that

$$Ae^{j\theta} = \bar{A}$$

if

$$A_1 = A\cos\theta \qquad \text{and} \qquad A_2 = A\sin\theta \qquad (5\text{-}11)$$

---

[*] Equation (5-10) can be regarded as a definition because the number *e* raised to an *imaginary* power has no meaning in the sense of a real exponent. It will be found that all operations which can be carried out with real exponentials will be true for Eq. (5-10). For example, consider the derivative of the left-hand side with respect to θ. Using the rule for real exponents, we get

$$de^{j\theta}/d\theta = je^{j\theta} = j(\cos\theta + j\sin\theta) = -\sin\theta + j\cos\theta$$

which is the same as the derivative of the right-hand side of Eq. (5-10). Also, with the aid of the identities for cos(x+y) and sin(x+y) it can be shown that

$$e^{jx}\,e^{jy} = \cos(x+y) + j\sin(x+y) = e^{j(x+y)}$$

In other words, when exponentials are multiplied, the exponents add, in similarity with real exponentials.

Assuming A_1 and A_2 were specified, these two equations can be solved for A and θ. Thus, squaring both equations and adding gives

$$A_1^2 + A_2^2 = A^2(\cos^2\theta + \sin^2\theta) = A^2$$

or

$$A = \sqrt{A_1^2 + A_2^2} \qquad (5\text{-}12)$$

Also, dividing one equation by the other gives

$$\frac{A_2}{A_1} = \frac{\sin\theta}{\cos\theta}$$

and so

$$\theta = \arctan\frac{A_2}{A_1} \qquad (5\text{-}13)$$

$Ae^{j\theta}$ is called the exponential form, whereas $A_1 + jA_2$ is called the rectangular form of $\bar{A}$. The latter terminology comes from Eqs. (5-11) which show that A_1 and A_2 can be viewed as the rectangular coordinates of a point whose position in a plane is specified in the polar coordinates A and θ. This is illustrated in Fig. 5-4. Since these rectangular coordinates are the components of a complex number, it is customary to label the axes "real" (meaning real part) and "imaginary" (meaning imaginary part). A pair of axes so labeled constitutes a complex plane. A is called the magnitude of $\bar{A}$, and θ its angle.

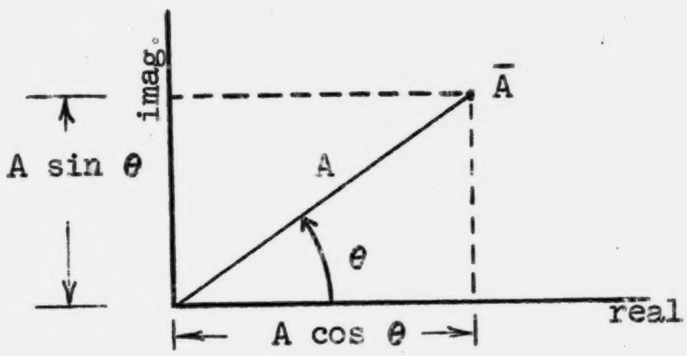


Figure 5-4.

The notation $Ae^{j\theta}$ is very convenient for analytical work, particularly when differentiation with respect to θ is involved. However, for specifying numerical values, like $2e^{j\pi/3}$, it is more convenient to write $2\underline{/60^\circ}$. This is called the polar form, since it merely involves specifying the polar coordinates of the point. (In the polar form, angles are usually specified in degrees.)

Observe that the exponential notation provides a simple interpretation for multiplication. Thus, if $\bar{A} = Ae^{j\theta_a}$ and $\bar{B} = Be^{j\theta_b}$ are two complex numbers, their product is

$$\overline{AB} = (AB)e^{j(\theta_a+\theta_b)}$$

In other words, magnitudes are multiplied and angles are added. Also note that in exponential notation the sign of the exponent is reversed in order to get a conjugate. Thus, if $\bar{A} = Ae^{j\theta}$, $\bar{A}^* = Ae^{-j\theta}$ (as can be seen by converting to the rectangular forms, since $\bar{A} = A \cos \theta + jA \sin \theta$ and $\bar{A}^* = A \cos \theta - jA \sin \theta$). In this notation, $\overline{AA}^* = A^2e^{j(\theta-\theta)} = A^2$, which is the same as $A_1^2 + A_2^2$.

5-3 Complex numbers applied to circuit theory.

Let us now consider how complex numbers can be used to represent a sinusoidal wave like

$$i = \sqrt{2}\, I \cos(\omega t + \alpha) \tag{5-14}$$

Since

$$Ie^{j(\omega t+\alpha)} = I \cos(\omega t+\alpha) + j \sin(\omega t+\alpha)$$

$$Ie^{-j(\omega t+\alpha)} = I \cos(\omega t+\alpha) - j \sin(\omega t+\alpha)$$

it follows that[+]

$$i = \sqrt{2}\, I \cos(\omega t+\alpha) = \sqrt{2}\, \frac{Ie^{j(\omega t+\alpha)} + Ie^{-j(\omega t+\alpha)}}{2} \tag{5-15}$$

The advantage to be gained by making this change depends upon defining an additional complex number

$$\bar{I} = Ie^{j\alpha} \tag{5-16}$$

[+] This is merely an application of the second of Eqs. (5-9).

Then, Eq. (5-15) can be written

$$i = \sqrt{2}\,\frac{Ie^{j\alpha}\,e^{j\omega t} + Ie^{-j\alpha}\,e^{-j\omega t}}{2}$$

$$= \sqrt{2}\,\frac{\bar{I}e^{j\omega t} + \bar{I}^{*}e^{-j\omega t}}{2} \qquad (5\text{-}17)$$

This formula is important. It shows that, given a sinusoid $i = \sqrt{2}\, I \cos(\omega t+\alpha)$, a complex number $\bar{I} = Ie^{j\alpha}$ can be constructed from the two parameters I and α which describe the wave, and that an expression for the sinusoid can be given in terms of that complex number. The importance of this idea cannot be overestimated. Conversely, if Eq. (5-17) is given, the properties of the wave are apparent merely from looking at the complex number $\bar{I}$. Observe that the quantity $\bar{I}^{*}$ is superfluous insofar as relationship of Eq. (5-17) to the wave is concerned, because $\bar{I}^{*}$ is completely determined by $\bar{I}$. To illustrate numerically,

$$\sqrt{2}\,\frac{3e^{j\pi/3}\,e^{j\omega t} + 3e^{-j\pi/3}\,e^{-j\omega t}}{2}$$

is an expression for $\sqrt{2}\,(3)\cos(\omega t+\pi/3)$. Likewise,

$$\sqrt{2}\,\frac{(2+j2)e^{j\omega t} + (2-j2)e^{-j\omega t}}{2}$$

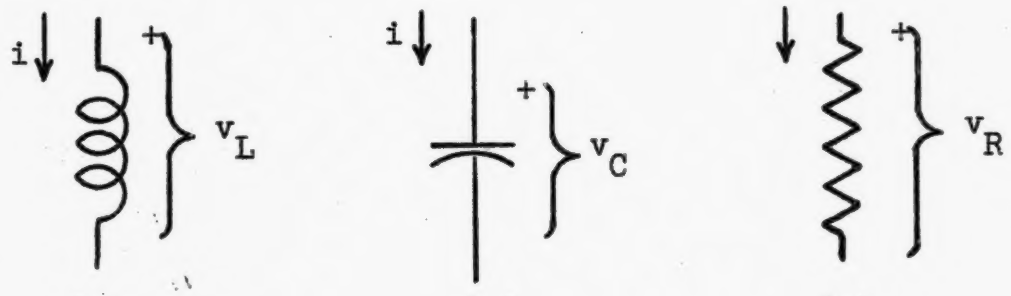
is an expression for $4\cos(\omega t+\pi/4)$, since $2 + j2 = 2\sqrt{2}\,e^{j\pi/4}$.

Referring back to Eq. (5-17), since $\bar{I}$ carries all the information (except frequency) necessary to construct the sinusoidal wave, it is given a special name. $\bar{I}$ is called the phasor symbol for the wave.[+] When a sinusoidal wave is expressed in complex form, the phasor symbol for that wave can always be recognized as the coefficient which multiplies $e^{j\omega t}$.

We shall now develop applications of Eq (5-17) to the steady state analysis of networks. Figure 5-5a shows an inductor, for which the voltage and current are related by

$$v_L = L\,\frac{di}{dt}$$

[+] In the older electrical engineering literature this is called a vector. However, calling $\bar{I}$ a vector is to be deprecated, because it does not have all the properties of a vector. Also, in some books it is called a sinor.



(a) (b) (c)

Figure 5-5.

Assume

$$i = \sqrt{2}\, I \cos(\omega t + \alpha) + I_{dc} \qquad (5\text{-}18)$$

where I_{dc} is a constant (d-c) term. The voltage v_L is obtained by differentiating the current and multiplying by L, giving

$$v_L = -\sqrt{2}\, \omega L\, I \sin(\omega t + \alpha) \qquad (5\text{-}19)$$

which is a sinusoidal wave. This can also be written as a cosine function by observing that $-\sin x = \cos(x + \pi/2)$, so that

$$v_L = \sqrt{2}\, \omega L\, I \cos(\omega t + \alpha + \frac{\pi}{2}) \qquad (5\text{-}20)$$

Particular note should be made of the fact that, even if the current contains a d-c component, the voltage will be a sinusoidal wave, because the derivative of a constant is zero. In many practical circuits, particularly those containing transistors or vacuum tubes, there will be d-c components. However, these can be calculated by applying the principles of d-c circuit analysis and, as illustrated by this example, the sinusoidal component of the current is realted to the sinusoidal voltage wave independently of the d-c component. Thus, in all subsequent consideration of inductor currents in this chapter we will deal only with a sinusoidal wave of current, recognizing that there may also be a d-c component, but that such a component will not be included in the expressions used for the inductor current.

Let us return to the sinusoidal component of current

$$i = \sqrt{2}\; I \cos(\omega t + \alpha)$$

and the voltage wave

$$v = \sqrt{2}\; \omega L\, I \cos(\omega t + \alpha + \tfrac{\pi}{2})$$

These can be written respectively in the complex notation

$$i = \sqrt{2}\, \frac{Ie^{j\alpha}\, e^{\,j\omega t} + Ie^{-j\alpha}\, e^{\,-j\omega t}}{2}$$

$$v = \sqrt{2}\, \frac{\omega L I e^{j(\alpha+\pi/2)}\, e^{j\omega t} + \omega L I e^{-j(\alpha+\pi/2)}\, e^{-j\omega t}}{2}$$

However, if we define the phasor $\bar{I} = Ie^{j\alpha}$, and recognize that $e^{j\pi/2} = j$ and $e^{-j\pi/2} = -j$, these become

$$i = \sqrt{2}\, \frac{\bar{I}e^{j\omega t} + \bar{I}^{*}e^{-j\omega t}}{2} \qquad (5\text{-}21)$$

$$v = \sqrt{2}\, \frac{j\omega L\bar{I}e^{j\omega t} - j\omega L\bar{I}e^{-j\omega t}}{2}$$

If we now define

$$\bar{V} = j\omega L\bar{I} \qquad (5\text{-}22)$$

the expression for v can be rewritten as

$$v = \sqrt{2}\, \frac{\bar{V}e^{j\omega t} + \bar{V}^{*}e^{-j\omega t}}{2} \qquad (5\text{-}23)$$

Equation (5-22) is important, for it gives the relationship between phasors for voltage and current in an inductor. It can be used in the form shown to find the parameters of the voltage wave if the current wave is given. If the voltage wave is specified, then $\bar{V}$ is known and $\bar{I}$ is found from the same relation, rewritten as

$$\bar{I} = \frac{\bar{V}}{j\omega L} = -j\,\frac{\bar{V}}{\omega L} \qquad (5\text{-}24)$$

A similar analysis applies to the capacitor shown in Fig. 5-5b. The voltage-current relationship is

$$i = C \frac{dv_C}{dt}$$

Suppose the voltage v_C is given by

$$v_C = \sqrt{2}\, V_C \cos(\omega t + \beta) + V_{dc} \qquad (5\text{-}25)$$

where V_{dc} represents a constant (d-c) component. Differentiating and multiplying by C gives

$$i = -\sqrt{2}\, \omega C V_C \sin(\omega t + \beta)$$

or

$$i = \sqrt{2}\, \omega C V_C \cos(\omega t + \beta + \frac{\pi}{2}) \qquad (5\text{-}26)$$

Just as it is possible for an inductor current to have a d-c component, so also there can be a d-c component of capacitor voltage. Steady state analysis for sinusoidal waves pertains to the relationship between the sinusoidal component of capacitor voltage and current; the d-c component of capacitor voltage can be found independently. Equations (5-25) and (5-26) show that V_{dc} has no effect on the relationship between the sinusoidal waves. Subsequently in this chapter, any symbol for capacitor voltage will be understood to refer to the sinusoidal component.

In complex number notation,

$$v_C = \sqrt{2}\, \frac{V_C\, e^{j\beta} e^{j\omega t} + V_C e^{-j\beta}\, e^{-j\omega t}}{2}$$

$$i = \sqrt{2}\, \frac{\omega C\, V_C e^{j(\beta+\pi/2)}\, e^{j\omega t} + \omega C\, V_C\, e^{-j(\beta+\pi/2)}\, e^{-j\omega t}}{2}$$

By defining the phasor $\bar{V}_C = V_C e^{j\beta}$ for the voltage, we can write these

$$v_C = 2\, \frac{\bar{V}_C\, e^{j\omega t} + \bar{V}_C{}^* e^{-j\omega t}}{2} \qquad (5\text{-}27)$$

$$i = \sqrt{2}\, \frac{\bar{I}\, e^{j\omega t} + \bar{I}^* e^{-j\omega t}}{2} \qquad (5\text{-}28)$$

where

$$\bar{I} = j\omega C\, \bar{V}_C \tag{5-29}$$

If the voltage wave is specified, $\bar{V}_C$ is known and Eq. (5-29) yields $\bar{I}$ and the current wave. If the current wave is specified, the voltage wave can be found from $\bar{V}_C$, where

$$\bar{V}_C = \frac{\bar{I}}{j\omega C} = - j\,\frac{\bar{I}}{\omega C} \tag{5-30}$$

The resistor case shown in Fig. 5-5c is quite simple, since no derivative is involved. If

$$i = \sqrt{2}\; I \cos(\omega t+\alpha)$$

then

$$v_R = \sqrt{2}\; IR \cos\,(\omega t+\alpha)$$

The respective complex number representations are

$$i = \sqrt{2}\;\frac{\bar{I}e^{j\omega t} + \bar{I}^{*}e^{-j\omega t}}{2}$$

and

$$v_R = \sqrt{2}\;\frac{R\bar{I}e^{j\omega t} + (R\bar{I})^{*}e^{-j\omega t}}{2}$$

Thus, if $\bar{I}$ and $\bar{V}_R$ are respectively the phasor symbols for i and v_R, then

$$\bar{V}_R = R\,\bar{I} \tag{5-31}$$

5-4 Phase Difference

In the previous section we found that, except for a resistor, current and voltage waves in a network element have different initial angles. To summarize, if in each case

$$i = \sqrt{2}\; I \cos(\omega t+\alpha) \tag{5-32}$$

then

$$\begin{aligned} v_R &= \sqrt{2}\,V_R \cos(\omega t+\alpha) & \text{(a)} \\ v_L &= \sqrt{2}\,V_L \cos(\omega t+\alpha+\tfrac{\pi}{2}) & \text{(b)} \\ v_C &= \sqrt{2}\,V_C \cos(\omega t+\alpha-\tfrac{\pi}{2}) & \text{(c)} \end{aligned} \qquad (5\text{-}33)$$

The corresponding waves are shown in Fig. 5-6.

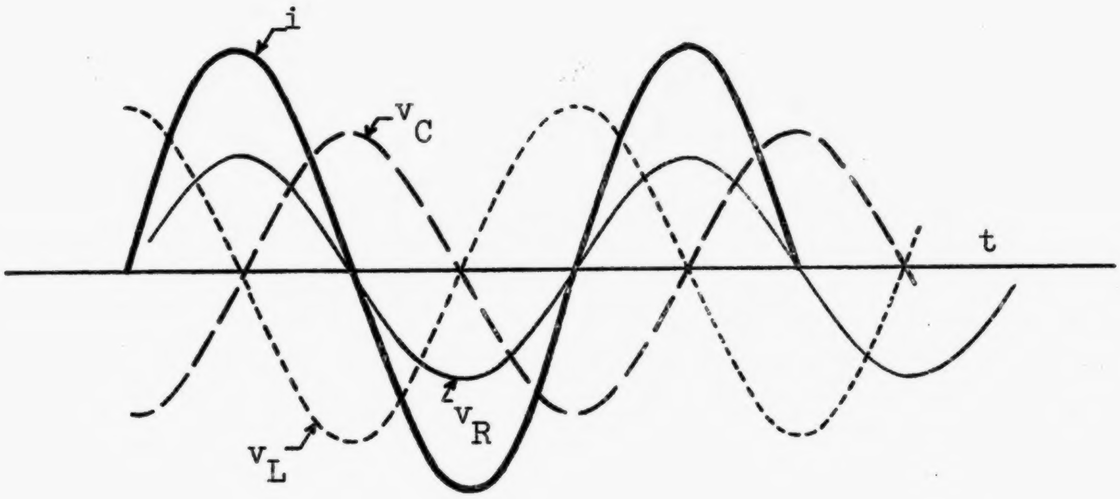


Figure 5-6.

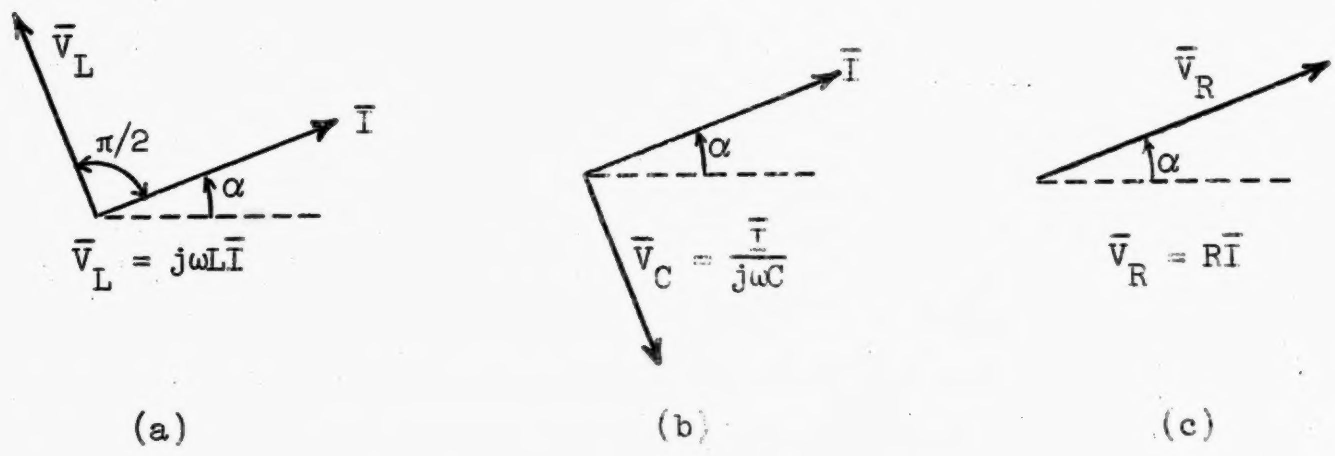
When two waves like i and v_R have the same initial angles, the waves are said to be in phase. If the initial angles are different, the peak values of the wave occur at different instants of time. For example, in Fig. 5-6 wave v_L reaches a peak value sooner than the i wave. It is said that v_L leads i. The amount of separation between peak values of two waves is measured as an angle (on the ωt scale), and this quantity is called the phase difference. In the case of v_L and i, the phase difference is $\pi/2$. Thus, a complete statement of the phase relationship is to say that v_L leads i by $\pi/2$ radians.

If we compare v_C with i, we find that v_C lags i (or i leads v_C) meaning the peak of v_C occurs later in time than the peak of i. In fact, v_C lags i by $\pi/2$ radians. When specifying phase difference, in addition to specifying the angle, it is necessary to state whether it is an angle of lead or lag. The possibility of specifying a phase difference either way causes some redundancy. For example, in the capacitor case, we can replace the statement that v_C lags i by $\pi/2$ radians by the equivalent statement that i leads v_C by $(-\pi/2)$ radians. Phase difference is an algebraic quantity for which the words "lead" or "lag" provide reference direction information. Since two reference directions are possible, there are two ways of specifying the phase difference.

Another kind of redundancy enters if angles greater than π are allowed. Thus, referring to Fig. 5-6, although the maximum labeled (2) of v_C occurs after maximum (1) of i, it is also true that maximum (3) of v_C occurs earlier than maximum (1). Thus, in addition to saying that v_C lags i by $\pi/2$ radians, it is correct to say that v_C leads i by $3\pi/2$ radians. This is merely an illustration of the fact that the phase difference can be increased or decreased by an integral multiple of 2π; this principle being due to the fact that maxima occur at 2π intervals in the ωt scale.

Phase differences can be obtained by calculating the difference between initial angles. To state this in such a way as to get the proper sign, if we wish to determine the angle by which one wave (designated as the first wave) leads another wave (designated as the second wave), the initial angle of the second wave should be subtracted from the initial angle of the first wave. For example, to find the angle by which v_L leads i, observe that there initial angles are respectively $\alpha + \pi/2$ and α. Thus, the angle by which v_L leads i is $\alpha + \pi/2 - \alpha = \pi/2$.

Phase difference, as well as magnitude relationships, can conveniently be shown on phasor diagrams, as shown in Fig. 5-7. These are graphical portrayals



(a) (b) (c)

Figure 5-7

of the three relations

$$\bar{V}_L = j\omega L\,\bar{I}$$

$$\bar{V}_C = \frac{\bar{I}}{j\omega C}$$

$$\bar{V}_R = R\bar{I}$$

where $\bar{I} = Ie^{j\alpha}$. We recall that v_L leads i by $\pi/2$ radians. This phase difference is evident as the angle between phasors $\bar{V}_L$ and $\bar{I}$, as indicated in Fig. 5-7a, and it is also evident that the phasor for the quantity which leads (v_L in this case) occupies a counter-clockwise position relative to the other one. This diagram also shows how it is possible to say that i leads v_L by $3\pi/2$, this being the counter-clockwise measured angle from $\bar{V}_L$ to $\bar{I}$.

Phase difference can be stated in terms of phasors. Thus, in addition to saying v_L leads i by $\pi/2$, it is also proper to say $\bar{V}_L$ leads $\bar{I}$ by $\pi/2$ radians. However, this latter statement is symbolic for the former.

5-5 Kirchhoff's Laws in Terms of Phasors

Consider a series circuit as shown in Fig. 5-8a, in which each box represents R, L or C, or a combination of these. The voltage across each box is sinusoidal and of the same frequency. That is, assume

$$v_1 = \sqrt{2}\,V_1 \cos(\omega t+\beta_1) \qquad \text{and} \qquad v_2 = \sqrt{2}\,V_2 \cos(\omega t+\beta_2)$$

In terms of the phasors

$$\bar{V}_1 = V_1 e^{j\beta_1} \qquad \text{and} \qquad \bar{V}_2 = V_2 e^{j\beta_2}$$

these voltages are

$$v_1 = \sqrt{2}\,\frac{\bar{V}_1 e^{j\omega t} + \bar{V}_1^{*} e^{-j\omega t}}{2} \qquad \text{and} \qquad v_2 = \sqrt{2}\,\frac{\bar{V}_2 e^{j\omega t} + \bar{V}_2^{*} e^{-j\omega t}}{2}$$

The Kvl equation for this figure is

$$v_T = v_1 + v_2 . \tag{5-34}$$

Substituting the expressions for v_1 and v_2 and collecting terms that multiply $e^{j\omega t}$ and $e^{-j\omega t}$, gives

$$v_T = \sqrt{2}\,\frac{(\bar{V}_1+\bar{V}_2)e^{j\omega t} + (\bar{V}_1+\bar{V}_2)^{*}e^{-j\omega t}}{2} \qquad (5\text{-}35)^{+}$$

The form of this shows that v_T is a sinusoidal wave of angular frequency ω, symbolically represented by

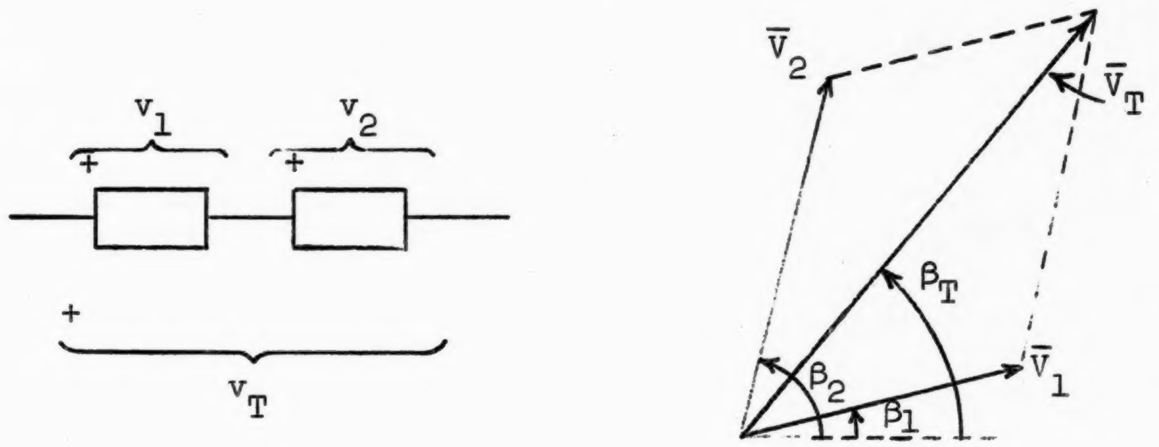
$$\bar{V}_T = \bar{V}_1 + \bar{V}_2 \qquad (5\text{-}36)$$

In other words, if we define the phasor

$$\bar{V}_T = V_T\, e^{j\beta_T}$$

the equation for the wave of v_T can also be written

$$v_T = \sqrt{2}\, V_T \cos(\omega t + \beta_T)$$



(a) (b)

Figure 5-8.

[+] In writing this equation, we have used the relation $\bar{V}_1^{\,*} + \bar{V}_2^{\,*} = (\bar{V}_1 + \bar{V}_2)^{*}$, which can readily be seen to be true by writing each phasor in rectangular form.

Equation (5-36) is an expression of the Kvl in terms of phasors. Of course, it is valid only when all quantities are sinusoidal. However, Eq. (5-35) shows that if v_1 and v_2 are each sinusoidal and of the same frequency, then v_T is also sinusoidal and of the same frequency. Equation (5-36) provides a means of finding V_T and β_T, which are the essential attributes of wave v_T. This can be done analytically, or graphically, as shown in Fig. 5-8b.

The Kirchhoff current law also can be written in terms of phasor quantities. Referring to Fig. 5-9a, if the currents i_1 and i_2 are sinusoidal and of the same frequency, namely,

$$i_1 = \sqrt{2}\, I_1 \cos(\omega t + \alpha_1) \quad \text{and} \quad i_2 = \sqrt{2}\, I_2 \cos(\omega t + \alpha_2)$$

they can be combined according to the Kcl equation

$$i_T = i_1 + i_2 \tag{5-37}$$

Again we shall perform the addition in terms of the complex notation, using the phasors $\bar{I}_1 = I_1 e^{j\alpha_1}$ and $\bar{I}_2 = I_2 e^{j\alpha_2}$ so that

$$i_1 = \sqrt{2}\,\frac{\bar{I}_1 e^{j\omega t} + \bar{I}_1^{*} e^{-j\omega t}}{2} \quad \text{and} \quad i_2 = \sqrt{2}\,\frac{\bar{I}_2 e^{j\omega t} + \bar{I}_2^{*} e^{-j\omega t}}{2}$$

The sum of these, after collecting terms, is

$$i_T = \sqrt{2}\,\frac{(\bar{I}_1+\bar{I}_2)e^{j\omega t} + (\bar{I}_1+\bar{I}_2)^{*} e^{-j\omega t}}{2} \tag{5-38}$$

which shows that i_T is a sinusoidal wave of angular frequency ω, and symbolically represented by a phasor $\bar{I}_T = I_T e^{j\alpha_T}$, where

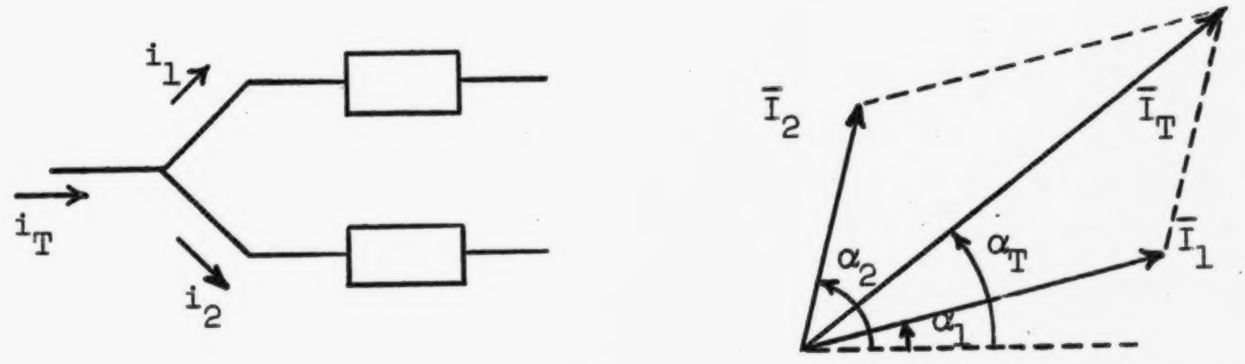
$$\bar{I}_T = \bar{I}_1 + \bar{I}_2 \tag{5-39}$$

In trigonometric form,

$$i_T = \sqrt{2}\, I_T \cos(\omega t + \alpha_T)$$

where I_T and α_T are obtainable from phasor $\bar{I}_T$ as indicated above.

Equation (5-39) is a valid expression for the Kcl for situations where i_1 an i_2 are both sinusoidal waves of the same frequency. We have shown that the sum o



(a) (b)

Figure 5-9.

two such sinusoidal current waves is also sinusoidal, and that the amplitude and initial angle of this wave can be obtained from Eq. (5-39). This equation can be used analytically; or a graphical interpretation like Fig. 5-9b can sometimes be useful.

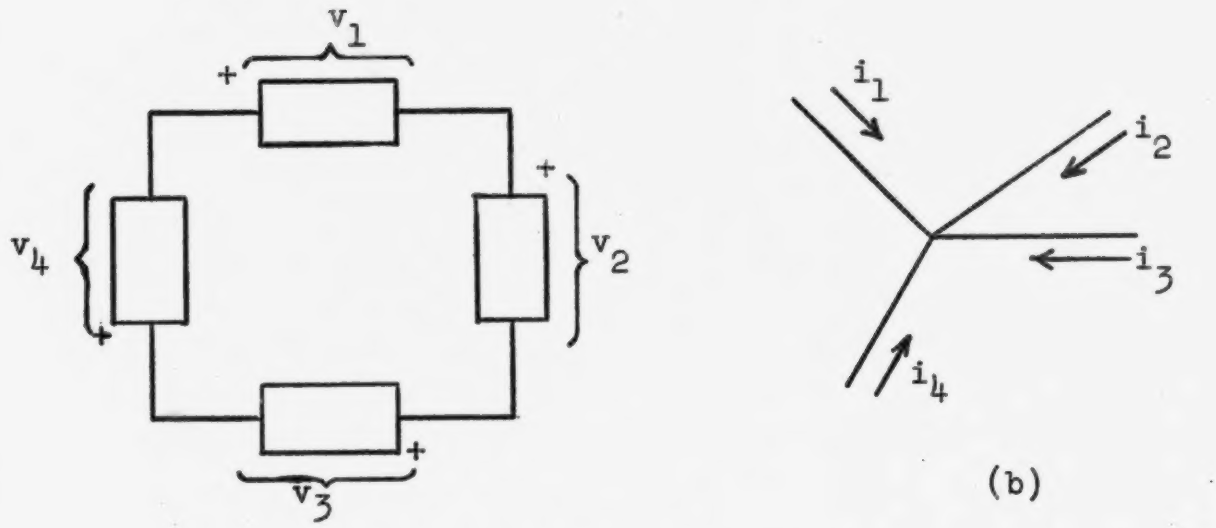
The proofs just given, involving the addition of two voltages or two currents, can be extended to yield Kvl and Kcl equations in phasor form for any number of voltages or currents, subject, of course, to the condition that all voltages and currents are sinusoids of the same frequency. Referring to Fig. 5-10a, one form of the Kvl equation is

$$-v_4 = v_1 + v_2 + v_3 \tag{5-40}$$

It is assumed that v_1, v_2, and v_3 are each sinusoidal and of the same frequency, and that they are respectively represented by phasors $\bar{V}_1$, $\bar{V}_2$, and $\bar{V}_3$. By the previous proof the sum of two of them, say $v_1 + v_2$, is also sinusoidal, and so these can be placed in parentheses, as a reminder that they represent one sinusoidal wave, thus,

$$-v_4 = (v_1+v_2) + v_3$$

But this is now reduced to the previous case, involving the addition of two sinusoidal waves $(v_1 + v_2)$ and v_3. Following this plan, we first apply Eq. (5-36) to show that the phasor for $(v_1 + v_2)$ is $(\bar{V}_1 + \bar{V}_2)$. The same equation applies again to show that the phasor for $(v_1 + v_2) + v_3$ is $(\bar{V}_1 + \bar{V}_2) + \bar{V}_3$. Thus,



(a) (b)

Figure 5-10.

from Eq. (5-40) we get

$$-\bar{V}_4 = (\bar{V}_1 + \bar{V}_2) + \bar{V}_3$$

or

$$0 = \bar{V}_1 + \bar{V}_2 + \bar{V}_3 + \bar{V}_4 \qquad (5\text{-}41)$$

A similar development will apply to the Kcl equations when currents are sinusoidal and of the same frequency. Referring to Fig. 5-10b, we have

$$-i_4 = i_1 + i_2 + i_3$$

$$= (i_1 + i_2) + i_3 \qquad (5\text{-}42)$$

If i_1, i_2, i_3, and i_4 have the respective phasors $\bar{I}_1$, $\bar{I}_2$, $\bar{I}_3$, and $\bar{I}_4$, Eq. (5-39) applies to show that $(i_1 + i_2)$ has the phasor $\bar{I}_1 + \bar{I}_2$, and applies again to show that $(i_1 + i_2) + i_3$ has the phasor $(\bar{I}_1 + \bar{I}_2) + \bar{I}_3$. Thus, Eq. (5-41) yields

$$-\bar{I}_4 = (\bar{I}_1 + \bar{I}_2) + \bar{I}_3$$

or

$$0 = \bar{I}_1 + \bar{I}_2 + \bar{I}_3 + \bar{I}_4 \qquad (5\text{-}43)$$

as the phasor expression of the Kcl equation.

The interpretation to be placed on the results of this section, culminating in Eqs. (5-41) and (5-43), is that Kvl and Kcl equations can be written in terms of phasor symbols in exactly the same way as for actual time-varying quantities. The arguments given for summing three sinusoidal quantities can be extended to any number. This statement is, of course, subject to the condition that all quantities involved shall be sinusoidally varying with time, and shall have the same frequency.

5-6 Analysis of Elementary Circuits

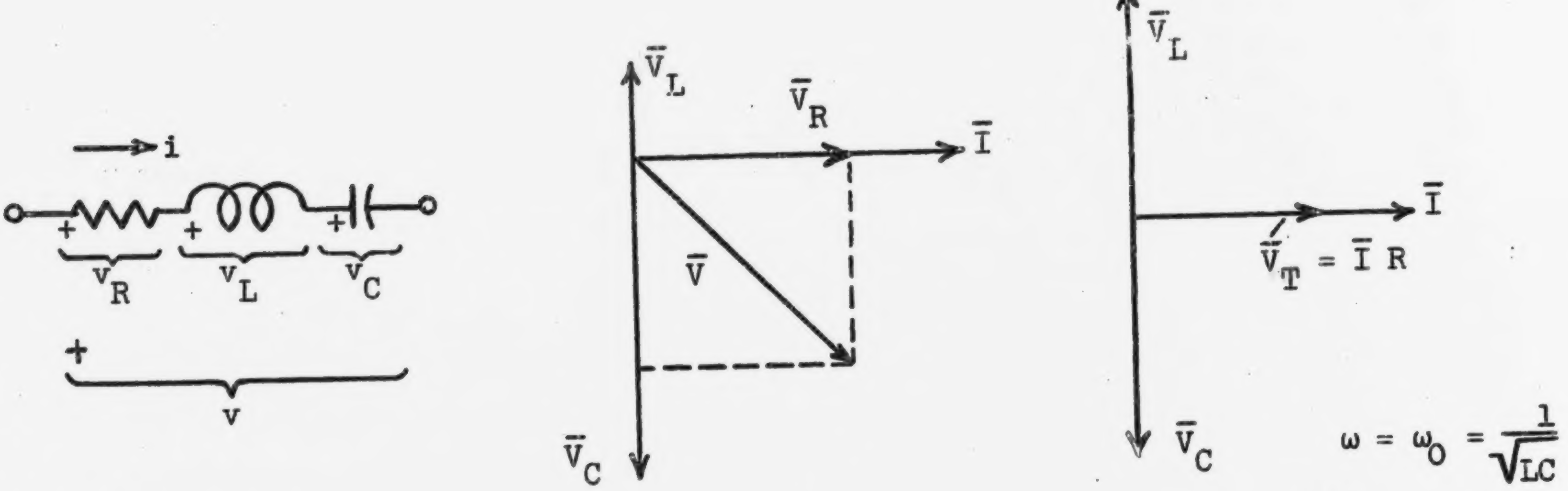
In the previous section it was shown that phasor quantities can be used to write Kirchhoff's laws, when waves are sinusoidal, and we have also established phasor relationships for R, L, and C elements. The next step is to use these principles to analyze some simple circuits.

As a first example, consider Fig. 5-11a, for which we assume that the current i is sinusoidal and represented by the phasor $\bar{I}$. From Eqs. (5-23), (5-30) and (5-31) we have, respectively, $\bar{V}_L = j\omega L\bar{I}$, $\bar{V}_C = \bar{I}/j\omega C$, and $\bar{V}_R = R\bar{I}$. Also, by virtue of the discussion in Sec. 5-5, Kvl can be applied in terms of phasors, giving

$$\bar{V} = \bar{V}_R + \bar{V}_L + \bar{V}_C \qquad (5\text{-}44)$$

or

$$\bar{V} = \left(R + j\omega L + \frac{1}{j\omega C}\right)\bar{I} = \left[R + j\left(\omega L - \frac{1}{\omega C}\right)\right]\bar{I} \qquad (5\text{-}45)$$



(a) (b) (c)

Figure 5-11.

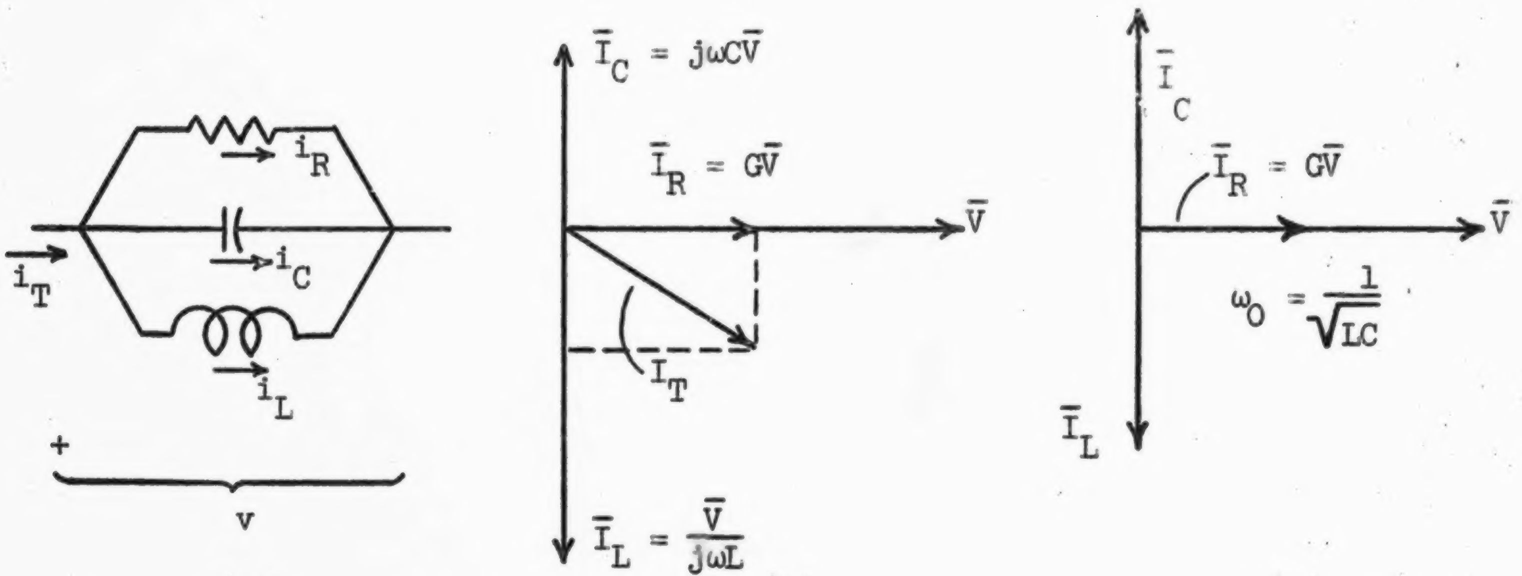
The three phasors on the right of Eq. (5-44), and also their sum, are shown in Fig. 5-11b, for a particular value of ω. For this case, the terminal voltage lags the current. However, it can be seen that for a sufficiently large value of ω, ωL will be larger than $1/\omega C$, so that $|\bar{V}_L|$ will be larger than $|\bar{V}_C|$. Then $\bar{V}$ will lie in the first quadrant, and will lead the current. The transition from voltage lag to voltage lead will occur when $\omega L = 1/\omega C$. This particular value of ω, which we shall designate as ω_0, given by

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{5-46}$$

is called the resonant angular frequency. The phasor diagram for this frequency is shown in Fig. 5-11c. In this case the phase difference between voltage and current is zero (the voltage and current are said to be in-phase).

This simple example has served to bring out several important aspects of alternating-current circuits: the application of complex numbers, resulting in Eq. (5-45) as a relationship between terminal voltage and current; the equivalent graphical analysis as displayed in the phasor diagram of Fig. 5-11b; and the notion that the relationship between current and voltage generally is dependent upon frequency (or ω), including the idea of a resonant frequency at which a circuit behaves like a pure resistance.

A similar analysis can be carried out for the parallel branch illustrated in Fig. 5-12a. Again referring to Eqs. (5-23), (5-30), and (5-31), we assume a



(a) (b) (c)

Figure 5-12.

sinusoidal voltage v, and have $\bar{I}_L = \bar{V}/j\omega L$, $\bar{I}_C = j\omega C\bar{V}$, and $\bar{I}_R = \bar{V}/R$. Also, from Kcl in phasor form, we have

$$\bar{I} = \bar{I}_R + \bar{I}_C + \bar{I}_L \qquad (5\text{-}47)$$

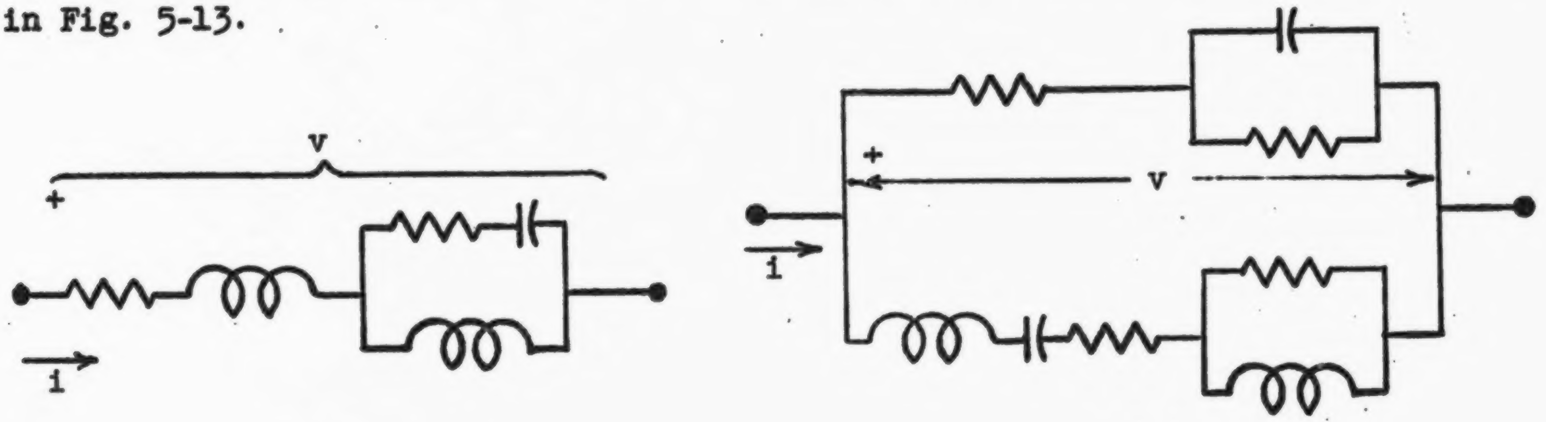
or

$$\bar{I} = (\frac{1}{R} + j\omega C + \frac{1}{j\omega L})\bar{V} = [\frac{1}{R}+j(\omega C - \frac{1}{\omega L})]\bar{V} \qquad (5\text{-}48)$$

Equation (5-48) is quite similar to Eq. (5-45), except that in the present case a voltage phasor is multiplied by a complex number to obtain the current phasor. A phasor diagram can be constructed in accordance with Eq. (5-47), as shown in Fig. 5-12b. Whether $\bar{I}$ lags or leads $\bar{V}$ depends on the frequency. In similarity with the series circuit, there is a resonant angular frequency, also given by Eq. (5-46), at which the voltage and current are in-phase.

5-7 Impedance and Admittance

Equations (5-45) and (5-48) apply to special cases of passive two-terminal networks. In general, these are interconnections of R, L and C elements, of any complexity that provides one or more current paths between two terminals. The word "passive" means that no sources can be included in any of the paths. The next step is to generalize to cases of greater complexity, such as the examples shown in Fig. 5-13.



(a) (b)

Figure 5-13.

First we note that for Fig. 5-11a we had the result

$$\bar{V} = \left[R + j(\omega L - \frac{1}{\omega C})\right]\bar{I} \qquad (a) \qquad (5\text{-}49)$$

which can also be written

$$\bar{I} = \left[\frac{1}{R + j(\omega L - \frac{1}{\omega C})}\right]\bar{V} \qquad \text{(b)} \qquad (5\text{-}49)$$

and for Fig. 5-12a the similar results are

$$\bar{I} = \left[\frac{1}{R} + j(\omega C - \frac{1}{\omega L})\right]\bar{V} \qquad \text{(a)}$$

which can also be written

$$(5\text{-}50)$$

$$\bar{V} = \left[\frac{1}{\frac{1}{R} + j(\omega C - \frac{1}{\omega L})}\right]\bar{I} \qquad \text{(b)}$$

In each case, one of these equations is superfluous (superfluous in the same sense that $v = Ri$ and $i = v/R$ are equivalent), but it is customary to adopt a notation and nomenclature which admits both forms. We shall now describe such a general notation. The above equations are examples of the following two forms

$$\bar{V} = \overline{ZI} \qquad (5\text{-}51)$$

$$\bar{I} = \overline{YV} \qquad (5\text{-}52)$$

in which $\bar{Z}$ and $\bar{Y}$ are given by the appropriate bracketed quantities. $\bar{Z}$ is called the impedance of the circuit, and $\bar{Y}$ the admittance. In view of Eqs. (5-51) and (5-52), it is obvious that

$$\bar{Y} = \frac{1}{\bar{Z}} \qquad (5\text{-}53)$$

so that if one of them is known, the other can easily be found. Although Eqs. (5-51) and (5-52) imply definitions for $\bar{Z}$ and $\bar{Y}$, it is worthwhile to define impedance and admittance explicitly as

$$\bar{Z} = \frac{\bar{V}}{\bar{I}} \quad \text{(a)} \quad \text{and} \quad \bar{V} = \frac{\bar{I}}{\bar{V}} \quad \text{(b)} \qquad (5\text{-}54)$$

In these definitions, and all the preceding formulas, it is understood that $\bar{V}$ is the phasor for the voltage across the terminals of the network in question, and that $\bar{I}$ is the phasor for the current in the terminals, with reference directions as portrayed in Fig. 5-13.

Since $\bar{Z}$ and $\bar{Y}$ are complex numbers, each can be written in rectangular form as

$$\bar{Z} = R + j\,X \tag{5-55}$$

$$\bar{Y} = G + j\,B \tag{5-56}$$

R and G carry the same names as in resistance networks, namely *resistance* and *conductance*, respectively. X is called the *reactance*, and B the *susceptance*. It should be noted that the similarity of the names of R and G with those used in resistive circuits should **not** be construed to mean that R (or G) refers to a single resistor. To take R as an example, if we consider the impedance of Fig. 5-11a, we find R of Eq. (5-55) is identical with the resistance R of the circuit. However, the impedance of Fig. 5-12a is

$$\bar{Z} = \frac{1}{\frac{1}{R} + j(\omega C - \frac{1}{\omega L})} = \frac{\frac{1}{R} - j(\omega C - \frac{1}{\omega L})}{(\frac{1}{R})^2 + (\omega C - \frac{1}{\omega L})^2}$$

In this case the R of Eq. (5-55) is

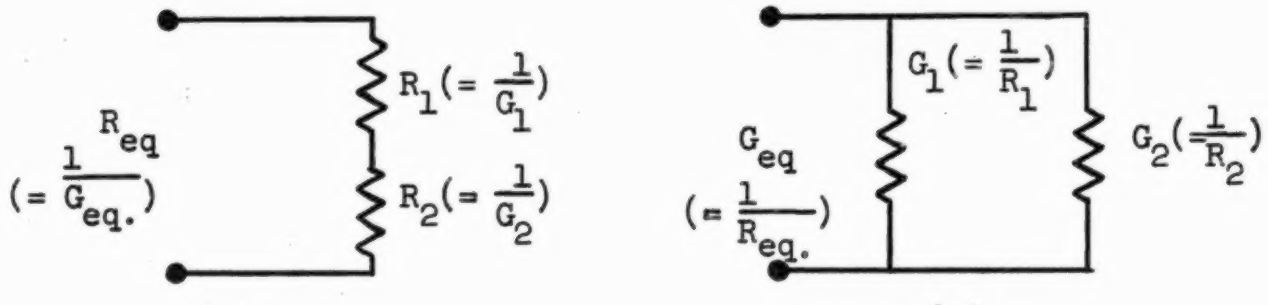
$$\frac{\frac{1}{R}}{(\frac{1}{R})^2 + (\omega C - \frac{1}{\omega L})^2}$$

which is not identifiable as a resistive circuit element of the original network. In Eqs. (5-55) and (5-56), R and G are to be regarded respectively as the real parts of $\bar{Z}$ and $\bar{Y}$, rather than as circuit elements.

We turn now to the question of how to find $\bar{Z}$ and $\bar{Y}$ for circuits of more general nature, such as the examples of Fig. 5-13. We shall do so by making note of similarities with resistance networks. It will be recalled that Kirchhoff's laws are used to prove the following relationships for the resistive networks of Fig. 5-14. -

$$\left.\begin{aligned} R_{eq} &= R_1 + R_2 \\ G_{eq} &= \frac{G_1 G_2}{G_1 + G_2} \end{aligned}\right\} \qquad \text{for Fig. 5-14a}$$

$$\left.\begin{aligned} G_{eq} &= G_1 + G_2 \\ R_{eq} &= \frac{R_1R_2}{R_1+R_2} \end{aligned}\right\} \quad \text{for Fig. 5-14b}$$


R_{eq}
$(=\frac{1}{G_{eq.}})$
$R_1(=\frac{1}{G_1})$
$R_2(=\frac{1}{G_2})$
G_{eq}
$(=\frac{1}{R_{eq.}})$
$G_1(=\frac{1}{R_1})$
$G_2(=\frac{1}{R_2})$

(a) (b)

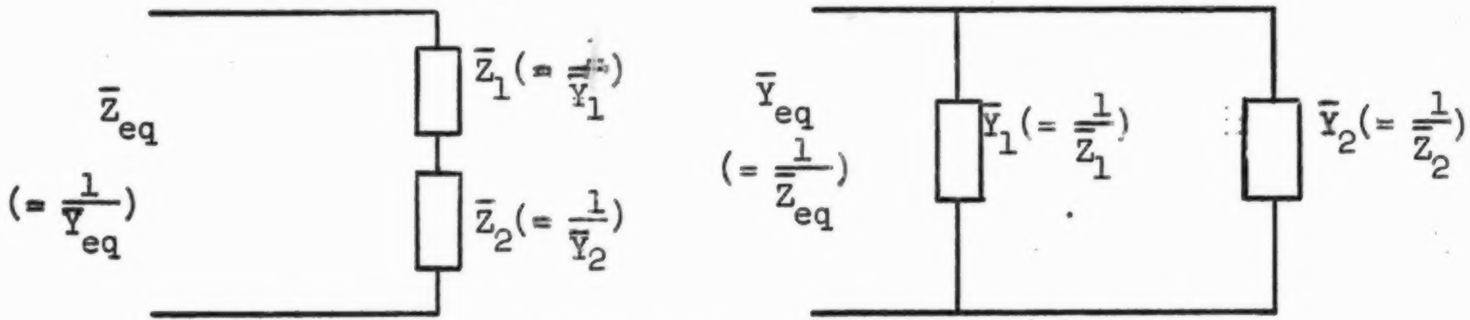
Figure 5-14.

These are the basic laws for combining resistances (or conductances) in series and parallel, which can be extended to more complicated cases merely by repetition.

By using Kirchhoff's laws in phasor notation, in proofs that are otherwise identical, it can be shown that impedances (or admittances) combine in similar ways. Referring to Fig. 5-15, the results are as follows:

$$\left.\begin{aligned} \bar{Z}_{eq} &= \bar{Z}_1 + \bar{Z}_2 \\ \bar{Y}_{eq} &= \frac{\bar{Y}_1\bar{Y}_2}{\bar{Y}_1+\bar{Y}_2} \end{aligned}\right\} \quad \text{for Fig. 5-14a} \qquad (5\text{-}57)$$

$$\left.\begin{aligned} \bar{Y}_{eq} &= \bar{Y}_1 + \bar{Y}_2 \\ \bar{Z}_{eq} &= \frac{\bar{Z}_1\bar{Z}_2}{\bar{Z}_1+\bar{Z}_2} \end{aligned}\right\} \quad \text{for Fig. 5-14b} \qquad (5\text{-}58)$$


$\bar{Z}_{eq}$
$(=\frac{1}{\bar{Y}_{eq}})$
$\bar{Z}_1(=\frac{1}{\bar{Y}_1})$
$\bar{Z}_2(=\frac{1}{\bar{Y}_2})$
$\bar{Y}_{eq}$
$(=\frac{1}{\bar{Z}_{eq}})$
$\bar{Y}_1(=\frac{1}{\bar{Z}_1})$
$\bar{Y}_2(=\frac{1}{\bar{Z}_2})$

(a) (b)

Figure 5-15.

These principles of combination can be used any number of times, starting with "rectangles" that represent branches that are simple enough to have known impedances or admittances (like Fig. 5-11a and Fig. 5-12a, for example, or special cases of these).

Although the procedure for combining impedances and admittances is algebraically identical with that for resistances (and conductances) the fact that complex numbers are involved in the present case means that numerical calculations are more laborious. Also, the translation from wave quantities (actual v and i) to phasor quantities and back must be included, a step which is not necessary for resistance networks. To illustrate these observations, the following two numerical examples are included:

Example 1: Referring to Fig. 5-16, assuming a terminal voltage $v_T = \sqrt{2}\,(75)\cos(3000t)$, the current i_T is to be found. The following numerical values will be used:

R_1 = 120 ohms $\quad$ R_2 = 2000 ohms ($G_2 = 5\text{x}10^{-4}$ mho)

L = 0.5 henry $\quad$ C = $0.2\text{x}10^{-6}$ farad

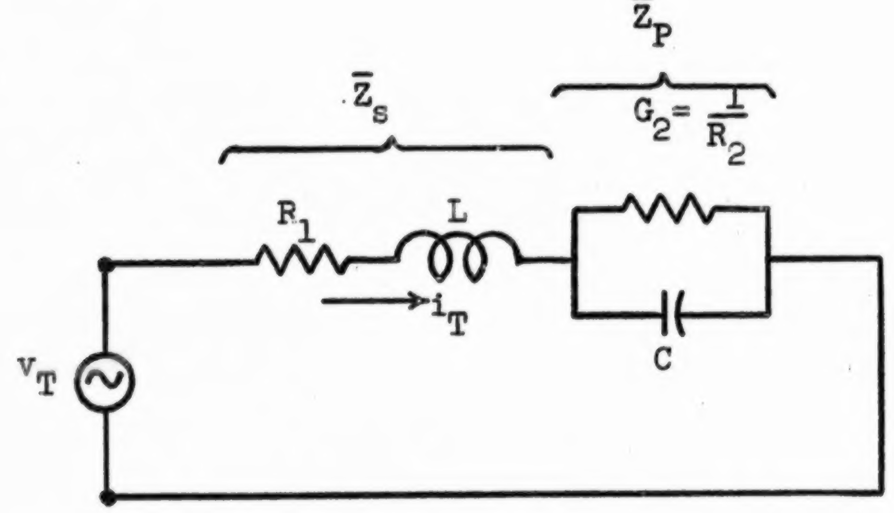


Figure 5-16.

Solution

Observe that the path between the source terminals can be viewed as two impedances (labeled $\overline{Z}_s$ and $\overline{Z}_p$) in series. Accordingly, these two impedances can be added, and the current I can be found from $\overline{I} = \overline{V}/\overline{Z}$. Furthermore, the portion labeled $\overline{Z}_s$ is a special case of Fig. 5-11a (with C replaced by a short circuit, meaning infinite C in the formula), and the portion labeled $\overline{Z}_p$ is a special case of Fig. 5-12a in which L is absent (infinite). Thus, we can proceed as follows:

Beginning with the parallel combination of G_2 and C, we note that it has an admittance

$$\frac{1}{\overline{Z}_p} = (5 + j6) \times 10^{-4}$$

where we have used $\omega C = 3000(0.2) \times 10^{-6} = 6 \times 10^{-4}$. The impedance of this parallel combination is therefore

$$\overline{Z}_p = \frac{10^4}{5+j6} = \frac{10^4}{7.80\underline{/50.2^\circ}} = 1282\underline{/-50.2^\circ}$$

In rectangular form this is[+]

$$\overline{Z}_p = 1282(\cos 50.2^\circ + j \sin 50.2^\circ)$$

$$\doteq 1282(.641 + j.768)$$

$$= 820 - j\,983$$

The other portion of the circuit has an impedance

$$\overline{Z}_s = 120 + j\,1500$$

[+] The rectangular form can be found directly, without going to the polar form, thus

$$\frac{10^4}{5+j6} = \frac{(5-j6)10^4}{(5+j6)(5-j6)} = \frac{(5-j6)10^4}{25 + 36} = \frac{5\text{x}10^4}{61} - j\,\frac{6\text{x}10^4}{61} = 820 - j\,983$$

However, in most cases, where the sum of the squares of the real and imaginary parts of the denominator cannot be found as easily as in this example, converting to polar coordinates is the simpler procedure. Many modern slide rules are constructed as to facilitate conversions between rectangular and polar coordinates.

where $\omega L = 3000(0.5) = 1500$. The total impedance is

$$\bar{Z}_T = \bar{Z}_p + \bar{Z}_s$$

$$= 940 + j\,517 = 1072 \underline{/28.8^o}$$

The voltage phasor is $\bar{V} = 75 + j0$, and so, having found $\bar{Z}$, the current phasor is

$$\bar{I}_T = \frac{\bar{V}}{\bar{Z}_T} = \frac{75 + j0}{1072\underline{/28.8^o}} = .0699 \underline{/-28.8^o}$$

Finally, the equation for the current (not its phasor) is

$$i_T = \sqrt{2}\,(.0699)\cos(3000t - 28.8^o)$$

Observe that the voltage leads the current and that the total impedance Z_T has a positive angle. Voltage leads current whenever impedance has a positive angle. However, impedance can also have a negative angle, in which case the voltage will lag the current. In this circuit, by carrying out the same calculations for $\omega = 200$, it will be found that

$$\bar{Z}_T = 120 + j\,100 + 1980 - j\,159$$

$$= 3000 - j\,59$$

which has a negative angle. In this case, the voltage will lag the current; by a small angle. [Note, in these statements, that voltage and current refer to voltage between the two terminals and current in one terminal (and out the other), using the customary reference markings for voltage and current

Example 2: In the circuit of Fig 5-17, assume that the current is known to be $i_T = \sqrt{2}\,(4)\cos\omega t$, where the frequency is 60 cps. Thus, $\omega = 2\pi(60) = 377$. The voltage v_T is required. The element values are

$R_1 = 100$ ohms ($G_1 = .01$ mho) $\qquad$ $R_2 = 150$ ohms

$L_1 = 1.5$ henry $\qquad$ $C_2 = 12 \times 10^{-6}$ farad

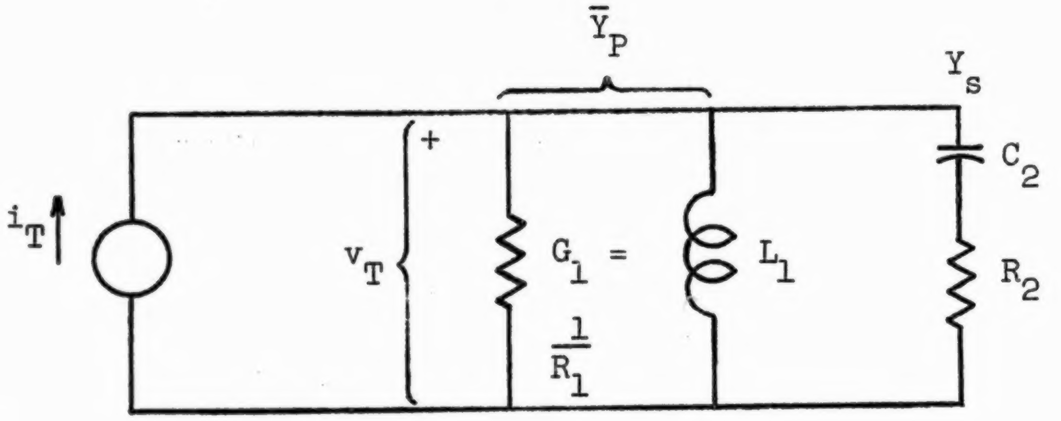


Figure 5-17.

Solution

This circuit has the general nature of a parallel connection, and can be viewed as having an equivalent admittance which is the sum of the two admittances $\overline{Y}_p$ and $\overline{Y}_s$. Thus, beginning with the R-C branch on the right, we write its impeda as a special case of Fig. 5-11a (with L = 0), giving

$$\overline{Z}_s = 150 - j\,\frac{10^6}{(377)(12)} = 150 - j\,221$$

$$= 268\ \underline{/-55.8^o}$$

and

$$\overline{Y}_s = \frac{1}{\overline{Z}_s} = 3.74 \times 10^{-3}\ \underline{/55.8^o} = .0021 + j\,.0031$$

For the parallel combination of L and R_1, the admittance is

$$\overline{Y}_p = .01 - j\,\frac{1}{(377)(0.5)} = .01 - j\,.00531$$

The total admittance is

$$\overline{Y}_T = \overline{Y}_s + \overline{Y}_p = .0021 + .01 + j(.0031 - .00531)$$

$$= .0121 - j\,.00221$$

The phasor $\overline{V}_T$ (symbolic for v_T) is

$$\bar{V}_T = \frac{\bar{I}_T}{\bar{Y}_T}$$

$$= \frac{4 + j0}{.0121 - j.00221} = \frac{4}{.01225\underline{/-10.4^\circ}} = 326\underline{/-10.4^\circ}$$

where $\bar{I}_T = 4 + j0$ is the phasor for the given current. Finally, the voltage wave is

$$v_T = \sqrt{2}\,(326)\cos(377t - 10.4^\circ)$$

5-8 Voltage and Current Source Equivalents

Following the general plan that is developing, whereby a-c circuits with R, L, and C elements can be treated like resistive circuits if phasor quantities are used, we now consider the equivalence of the two networks in Fig. 5-18.

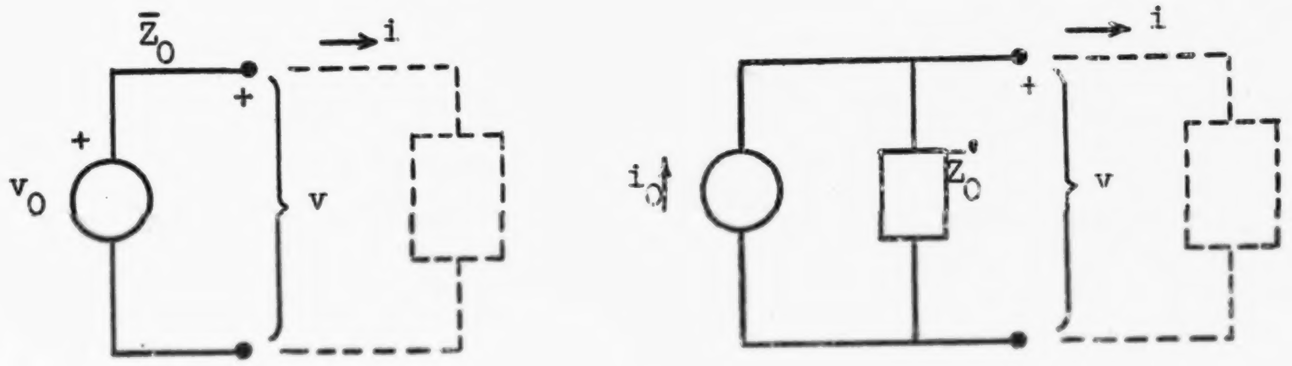


Figure 5-18

The ideal voltage source v_0 maintains a sinusoidal voltage

$$v_0 = \sqrt{2}\,V_0 \cos(\omega t + \beta_0)$$

under all conditions, and the ideal current source maintains a sinusoidal current

$$i_0 = \sqrt{2}\,I_0 \cos(\omega t + \alpha_0)$$

under all conditions. These sources are respectively symbolically represented by phasors $\bar{V}_0 = V_0 e^{j\beta_0}$ and $\bar{I}_0 = I_0 e^{j\alpha_0}$. The dotted rectangles shown in each case represent identical two-terminal networks consisting of R, L, and C elements and possibly sinusoidal sources, also of angular frequency ω. The presence of these external branches permits a current i to flow, and by restricting what can be in these branches we ensure that v and i will be sinusoidal (of angular frequency ω) and hence can respectively be represented by phasors $\bar{V}$ and $\bar{I}$.

Just as in the resistive network case, respectively for (a) and (b) of Fig. 5-18, we can write

$$\begin{aligned} \bar{V} &= \bar{V}_0 - \bar{Z}_0\bar{I} & \text{(a)} \\ \bar{I} &= \bar{I}_0 - \frac{1}{\bar{Z}_0'}\bar{V} & \text{(b)} \end{aligned} \qquad (5\text{-}59)$$

The second of these can be written

$$\bar{V} = \bar{Z}_0'\bar{I}_0 - \bar{Z}_0'\bar{I} \qquad (5\text{-}60)$$

In order for the two networks of Fig. 5-18 to be equivalent, they must be equivalent for all $\bar{V}$ and $\bar{I}$ (i.e. $\bar{V}$ and $\bar{I}$ must be the same for both). Thus, equating Eqs. (5-59) and (5-60), to ensure that $\bar{V}$ will be the same, we have

$$\bar{V}_0 - \bar{Z}_0\bar{I} = \bar{Z}_0\bar{I}_0 - \bar{Z}_0'\bar{I}$$

or

$$(\bar{V}_0 - \bar{Z}_0\bar{I}_0) + (\bar{Z}_0' - \bar{Z}_0)\bar{I} = 0$$

This equation must be true for all values of $\bar{I}$, and so each term must be independently zero, giving

$$\bar{V}_0 = \bar{Z}_0\bar{I}_0 \qquad (5\text{-}61)$$

as the relationship between the sources, and

$$\bar{Z}_0' = \bar{Z}_0 \qquad (5\text{-}62)$$

as the relationship between the impedances.

The equivalence between these two networks can be used to reduce circuit complexity in the same way as in resistance networks. However, it must be kept in mind that the present case is more restrictive in that the sources must be sinusoidal whereas in resistive networks any manner of time variation is permitted.

This is the "price we pay" for using phasor symbolism,[+] but many practical problems fall into this category.

There is one other important exception to the similarities between the solution of resistance networks and the steady state solution of RLC networks. In the case of resistance networks, it is possible to include ideal diodes, or piecewise linear equivalents, with due regard to the fact that the circuit changes whenever the diode current goes through zero. Phasors cannot be used for sinusoidal waves if diodes are included in a circuit, unless the diodes are biased sufficiently so that the current never reaches zero. Otherwise, if the diode goes from a conducting to a non-conducting condition during the swing of the signal, certain voltages and currents will become non-sinusoidal. This violates the conditions under which phasors can be used. Other methods of analysis must be used in such cases. There is no intent here to explain how to solve such cases, only to point out a limitation on the extension of resistance networks methods to sinusoidal steady state methods of analysis.

5-9 Thévenin and Norton Theorems

For any R, L, C network having only two terminals, such as the box on the left of Fig. 5-19, subject to the conditions that all sources within the box are of the same frequency, an equation of the form

$$\bar{V} = \bar{V}_0 - \bar{Z}_0 \bar{I} \qquad (5\text{-}63)$$

will be found as the relationship between $\bar{V}$ and $\bar{I}$. In this case we do not find a source V_0 or an impedance $\bar{Z}_0$. These are quantities that might be derived from

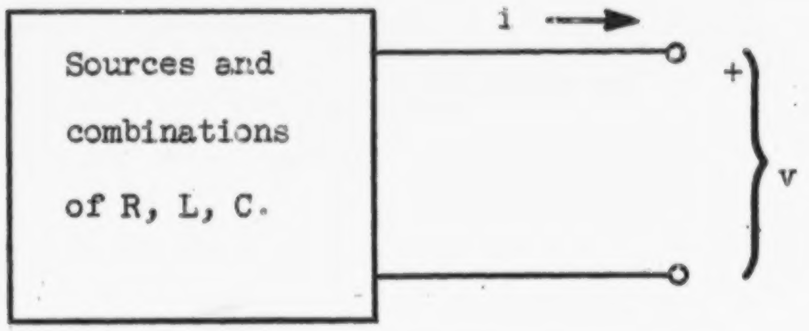


Figure 5-19.

[+] This is not to say that the general case of any type of source cannot be handled for R, L, C circuits, but to do this requires mathematical methods beyond those used in this text.

an analysis of the actual network (analysis by the solution of Kirchhoff equations, or perhaps by a succession of applications of the source equivalents). This means that $\overline{V}_0$ and $\overline{Z}_0$ are functions of the elements within the box, and how they are connected. A physical interpretation of them is possible, as follows: From Eq. (5-63), it is seen that $\overline{V} = \overline{V}_0$ when $\overline{I} = 0$. Thus, $\overline{V}_0$ is the open circuit value of $\overline{V}$. $\overline{V}_0$ is due to the internal sources, and will be zero if these sources are made zero. This fact provides an interpretation of $\overline{Z}_0$ as $-\overline{V}/\overline{I}$, when $\overline{V}_0 = 0$, (i.e., all internal sources are zero). Observe that $-\overline{I}$ is the current flowing into the + terminal, so that it is evident that $-\overline{V}/\overline{I}$ is the impedance of the network when all sources are zero.

> This equivalence can be stated in the following form: A network having one pair of terminals and consisting of interconnections of R, L and C elements, and voltage and current sources, all of which are sinusoidal with the same frequency, is equivalent at the pair of terminals to a voltage source having phasor $\overline{V}_0$ in series with an impedance $\overline{Z}_0$. $\overline{V}_0$ is the value of the terminal voltage phasor when the terminals are open-circuited; and Z_0 is the impedance at the terminals of the network when all the sources are deactivated.

Having already established the equivalence of circuits (a) and (b) of Fig. 5-18, it is evident that either of these circuits is equivalent to the original one, if $\overline{V}_0$, $\overline{Z}_0$, and $\overline{I}_0 = \overline{V}_0/\overline{Z}_0$ are interpreted as described above.

Figure 5-18a is called the Thévenin equivalent, and the theorem which states this equivalent is called Thévenin's Theorem. Also Fig. 5-18b is the Norton equivalent, and a statement of the equivalence is Norton's Theorem.

5-10 Methods of Network Analysis

Concerning steady state sinusoidal waves all of the same angular frequency ω, we can summarize the following:

1) Waves like $i = \sqrt{2}\, I \cos(\omega t+\alpha)$ and $v = \sqrt{2}\, V \cos(\omega t+\beta)$ can be treated symbolically in terms of the phasors $\bar{I} = Ie^{j\alpha}$ and $\bar{V} = Ve^{j\beta}$. These phasors do not equal i and v, however.
2) Kirchhoff's voltage and current equations can be written in terms of phasor symbols, to obtain the symbolic equivalent of writing the corresponding equations for the sinusoidal waves.
3) For individual R, L, or C elements, the voltage and current phasors are related by

$$\bar{V} = R\,\bar{I}$$

$$\bar{V} = j\omega L\,\bar{I}$$

$$\bar{V} = \frac{1}{j\omega C}$$

4) For any two-terminal network (branch) of R, L, and C elements the relationship between $\bar{V}$ and $\bar{I}$ can be written either as $\bar{V} = \overline{ZI}$ or $\bar{I} = \overline{YV}$, and $\bar{Z}$ and $\bar{Y}$ can be found by methods similar to the series and parallel combination of resistances in resistance networks. Voltage and current divider principles also apply.
5) Thévenin and Norton theorems apply, using phasor quantities.

We shall use a practical example to illustrate further generalizations of methods first introduced in the chapter on resistance networks. In so doing, the important concepts of loop and node equations will be introduced. The fact that these are included only in an example should not be construed as a depreciation of their importance. They are not given a specific theoretical discussion because by now it seems that the close similarity with resistance networks should be accepted, so that it will be evident that resistance network principles will apply, with $\bar{V} = \overline{ZI}$ replacing $v = Ri$, and $\bar{I} = \overline{YV}$ replacing $i = v/R$.

The example is the low-pass filter network shown in Fig. 5-20. Such a network gives preferential treatment to waves of low frequencies. This example will be

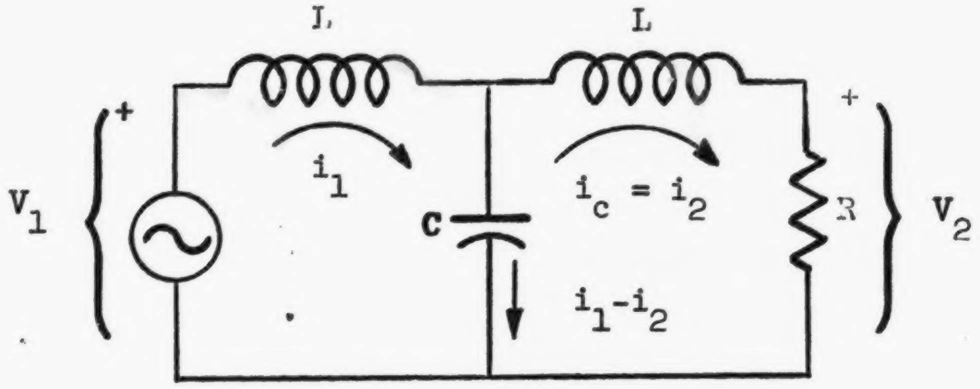


Figure 5-20.

used to illustrate various methods of determining the output voltage v_2, when the input voltage v_1 is specified. The input voltage will be assumed to be

$$v_1 = \sqrt{2}\, V_1 \cos \omega t$$

so that its phasor is $V_1 e^{j0} = V_1 + j0$.

(a) Loop Analysis

Loop currents i_1 and i_2 (with phasors $\bar{I}_1$ and $\bar{I}_2$) will be used, so that the current i_c in the C element will be $i_1 - i_2$. Two Kvl equations will be obtained. For the left and right hand loops, respectively, these are

$$\begin{aligned} j\omega L\, \bar{I}_1 + \frac{1}{j\omega C}(\bar{I}_1 - \bar{I}_2) &= \bar{V}_1 \quad \text{(a)} \\ -\frac{1}{j\omega C}(\bar{I}_1 - \bar{I}_2) + (R + j\omega L)\bar{I}_2 &= 0 \quad \text{(b)} \end{aligned} \qquad (5\text{-}64)$$

The second of these equations can be solved for $\bar{I}_1$, to give

$$\begin{aligned} \bar{I}_1 &= j\omega C\, [R + j\omega L + \frac{1}{j\omega C}]\, \bar{I}_2 \\ &= (1 - \omega^2 LC + j\omega RC)\, \bar{I}_2 \end{aligned}$$

This can be substituted in Eq. (5-64a), giving

$$j(\omega L - \frac{1}{\omega C})(1 - \omega^2 LC + j\omega RC)\bar{I}_2 + \frac{j}{\omega C}\bar{I}_2 = \bar{V}_1$$

$$\left\{(\frac{1}{\omega C} - \omega L)(\omega RC) + j[(\omega L - \frac{1}{\omega C})(1 - \omega^2 LC) + \frac{1}{\omega C}]\right\}\bar{I}_2 = \bar{V}_1$$

$$[(1 - \omega^2 LC)R + j\omega L(2 - \omega^2 LC)]\ \bar{I}_2 = \bar{V}_1$$

and, finally, the voltage phasor ($\bar{V}_2 = \overline{RI_2}$) is

$$\bar{V}_2 = \left[\frac{1}{(1-\omega^2 LC) + j\ \frac{\omega L}{R}(2-\omega^2 LC)}\right]\bar{V}_1 \qquad (5\text{-}65)$$

Before interpreting this result, we shall obtain Eq. (5-65) by two other methods.

(b) Node Analysis

The circuit under consideration is redrawn in Fig. 5-21, showing a node (3) in the center and corresponding voltage v_3. First let us sum the currents entering node 3. This sum must be zero, and so we get

$$\frac{1}{j\omega L}(\bar{V}_1 - V_3) - j\omega C\ V_3 + \frac{1}{j\omega L}(\bar{V}_2 - V_3) = 0$$

or

$$\frac{1}{j\omega L}\bar{V}_1 + \frac{1}{j\omega L}\bar{V}_2 - (\frac{2}{j\omega L} + j\omega C)\bar{V}_3 = 0$$

Multiplying this through by $j\omega L$ gives

$$\bar{V}_1 + \bar{V}_2 - (2 - \omega^2 LC)\bar{V}_3 = 0 \qquad (5\text{-}66)$$

This equation involves two unknowns ($\bar{V}_2$ and $\bar{V}_3$) since $\bar{V}_1$ is known. Thus, one other equation is needed. This equation is obtained from node 2. The current from (3) to (2) is

$$\frac{1}{j\omega L}(\bar{V}_3 - \bar{V}_2)$$

and the current leaving node 2, through R, is $\bar{V}_2/R$. We equate these, to give

$$\frac{V_2}{R} = \frac{1}{j\omega L}(\bar{V}_3 - \bar{V}_2)$$

or

$$\bar{V}_3 = (\frac{j\omega L}{R} + 1)\,\bar{V}_2 \qquad (5\text{-}67)$$

Substituting this expression for $\bar{V}_3$ into Eq. (5-66) yields

$$\bar{V}_1 - \left[(2 - \omega^2 LC)(\frac{R+j\omega L}{R}) - 1\right]\bar{V}_2 = 0$$

The quantity within the brackets simplifies to

$$(1 - \omega^2 LC) + \frac{j\omega L}{R}(2 - \omega 2LC)$$

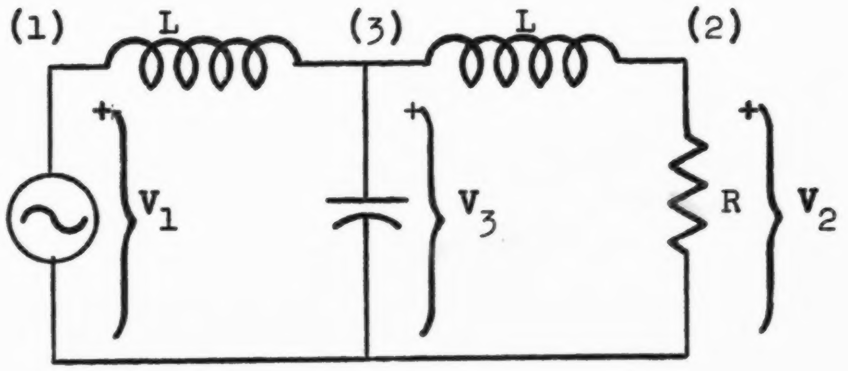


Figure 5-21.

and so the result is

$$\bar{V}_2 = \frac{\bar{V}_1}{(1-\omega^2 LC) + \frac{j\omega L}{R}(2-\omega^2 LC)}$$

which is in agreement with Eq. (5-65).

(c) Thévenin's Theorem

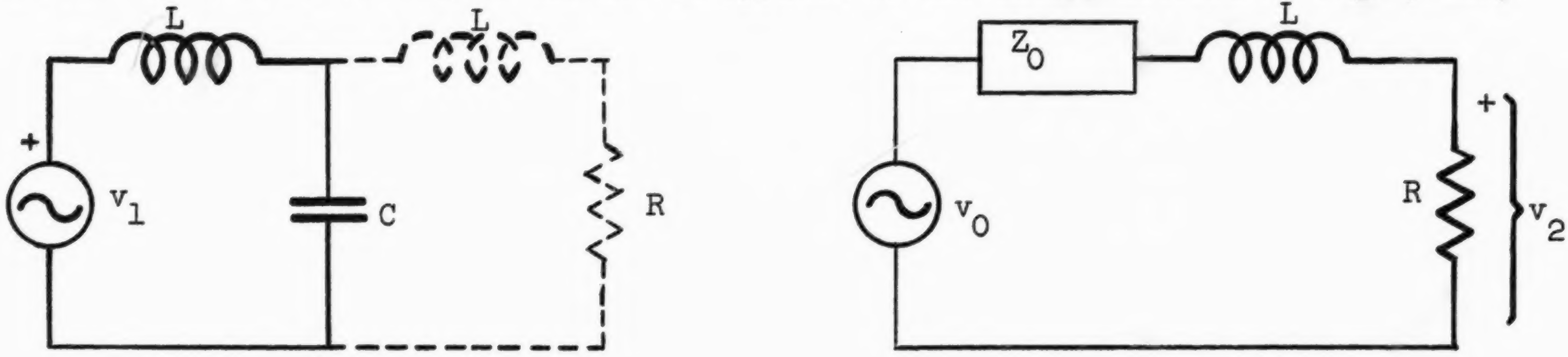
The same network can be solved by Thévenin's theorem by using the equivalence shown in Fig. 5-22. The section shown in solid lines at (a) is replaced by the circuit shown at (b). To accomplish this, we first refer to Fig. 5-22a, and observe that the open circuit value of $\bar{V}_{30}$ (which we shall designate as $\bar{V}_0$) can be obtained from the voltage divider principle, as

$$\bar{V}_0 = \frac{\frac{1}{j\omega C}\bar{V}_1}{j\omega L + \frac{1}{j\omega C}} = \frac{\bar{V}_1}{1 - \omega^2 LC} \qquad (5\text{-}68)$$

The impedance $\bar{Z}_0$ is obtained from this figure by deactivating the source (replacing it by a short circuit) and computing the impedance between the terminals of the resulting network. This gives a network of L and C in parallel, for which the impedance is

$$\bar{Z}_0 = \frac{j\omega L(\frac{1}{j\omega C})}{j\omega L + \frac{1}{j\omega C}} = \frac{j\omega L}{1 - \omega^2 LC} \qquad (5\text{-}69)$$

The voltage divider principle can now be applied to the entire network of Fig. 5-21b,



(a)

(b)

Figure 5-22.

giving

$$\bar{V}_2 = \frac{R\left(\frac{\bar{V}_1}{1-\omega^2 LC}\right)}{\frac{j\omega L}{1-\omega^2 LC} + j\omega L + R}$$

$$= \frac{\bar{V}_1}{1 - \omega^2 LC + j\frac{\omega L}{R}(2-\omega^2 LC)}$$

in agreement with Eq. (5-65).

We have now obtained the same formula for $\bar{V}_2$ by three different methods. In order to avoid leaving this as an abstract formula, let us consider its significance for some numerical cases. In so doing, we will not only add meaning to the formula, we shall be able to demonstrate more clearly what is meant by saying that the circuit in question is a low-pass filter.

The transmission at various frequencies can be determined from Eq. (5-65) by rewriting it as

$$\frac{\bar{V}_2}{\bar{V}_1} = \frac{1}{1 - \omega^2 LC + j\frac{\omega L}{R}(2-\omega^2 LC)} \qquad (5\text{-}70)$$

and assuming various numerical values for ω, from which numerical results can be obtained. Observe that there are only two combinations of parameters which enter into this expression, the product LC and the ratio L/R. Let these have the values $LC = 10^{-6}$ sec^2 and $L/R = 10^{-3}$ sec.

The following table can be constructed:

Table

ω	$\omega^2 LC$	$1-\omega^2 LC$	$2-\omega^2 LC$	$.001\omega(2-\omega^2 LC)$	$1-\omega^2 LC + j\frac{\omega L}{R}(2-\omega^2 LC)$	$\bar{V}_2/\bar{V}_1$
100	.01	.99	.99	.20	$1.00+j.20=1.02\underline{/11^\circ}$	$.98\underline{/-11^\circ}$
300	.09	.91	1.91	.57	$.91+j.57=1.0=\underline{/32^\circ}$	$.93\underline{/-32^\circ}$
500	.25	.75	1.75	.88	$.75+j.88=1.13\underline{/50^\circ}$	$.88\underline{/-50^\circ}$
800	.64	.36	1.36	1.09	$.36+j1.09=1.15\underline{/72^\circ}$	$.87\underline{/-72^\circ}$
1000	1.00	0	1.00	1.00	$0+j=1\underline{/90^\circ}$	$1\underline{/-90^\circ}$
1200	1.44	- .44	.56	.67	$-.44+j.67=.80\underline{/123^\circ}$	$1.25\underline{/-123^\circ}$
1300	1.69	- .69	.31	.42	$-.69+j.40=.80\underline{/150^\circ}$	$1.25\underline{/-150^\circ}$
1414	2.00	-1.00	0	0	$-1+j0 = 1\underline{/180^\circ}$	$1\underline{/-180^\circ}$
1600	2.56	-1.56	- .56	-.90	$-1.56-j.90=1.80\underline{/210^\circ}$	$.56\underline{/-210^\circ}$
2000	4	-3	-2	- 4.0	$-3.0-j4.0 = 5.0\underline{/233^\circ}$	$.20\underline{/-233^\circ}$
3000	9	-8	-7	-21	$-8.0-j21=22.4\underline{/249^\circ}$	$.05\underline{/-249^\circ}$
4000	16	-15	-14	-56	$-5.0-j56=56\underline{/265^\circ}$	$.02\underline{/-265^\circ}$

The most direct way to portray the complex quantity $\bar{V}_2/\bar{V}_1$ as a function of ω, is to plot the results carried in the right-hand column of the above tabulation, as in Fig. 5-23a. Imagine a unit voltage v_1 (rms value = 1) so that its wave is described by

$$v_1 = \sqrt{2}\cos\omega t \quad ; \qquad \text{and} \qquad \bar{V}_1 = 1 + j0$$

where ω can have any value in the range covered by the table and Fig. 5-23a.

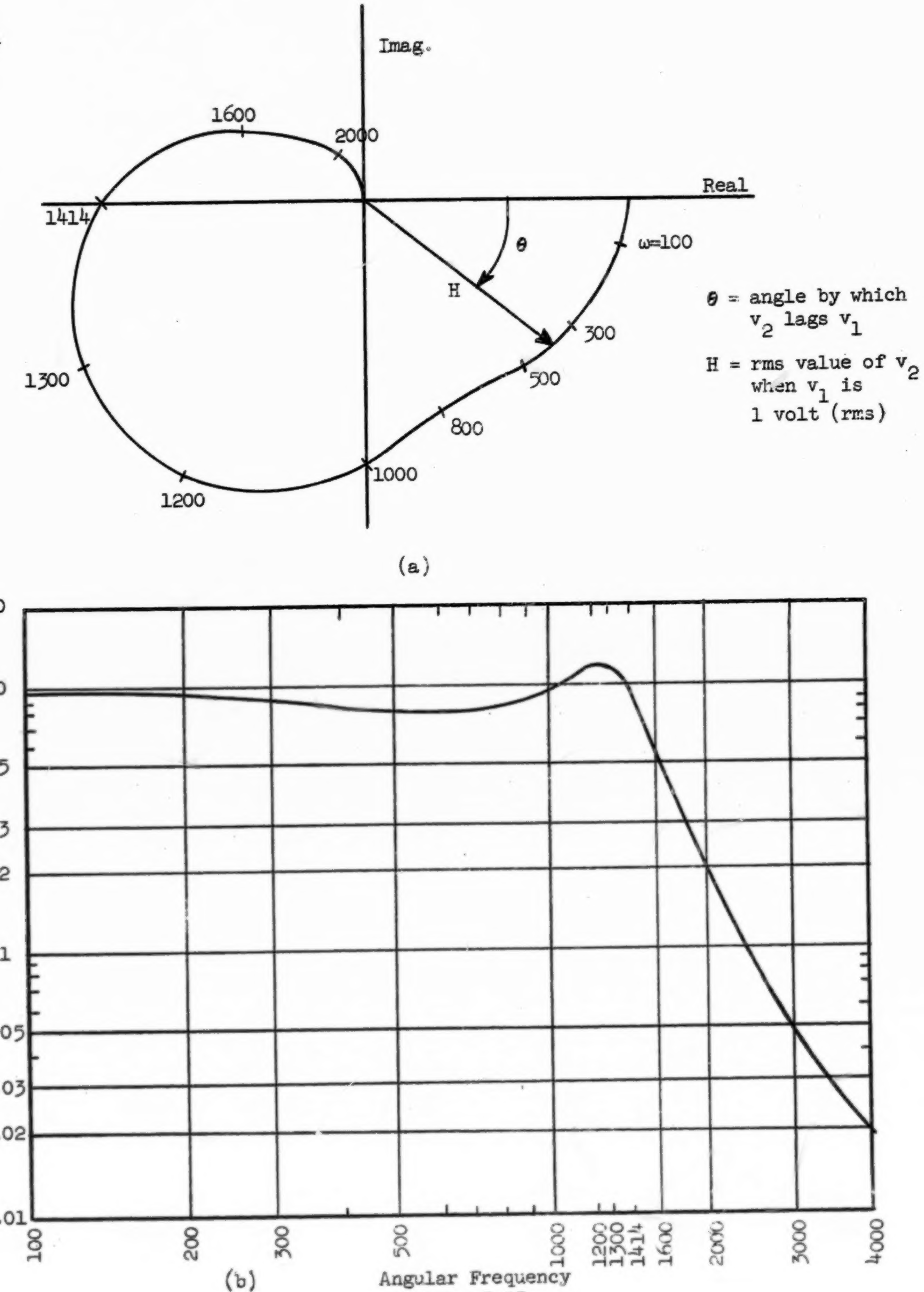

Imag.
Real
1600
2000
1414
ω=100
θ
H
300
500
800
1000
1200
1300
θ = angle by which v_2 lags v_1
H = rms value of v_2 when v_1 is 1 volt (rms)
(a)
2.0
1.0
.5
.3
H .2
.1
.05
.03
.02
.01
100
200
300
500
1000
1200
1300
1414
1600
2000
3000
4000
(b)
Angular Frequency

Figure 5-23.

At any frequency $\bar{V}_2/\bar{V}_1 = He^{-j\theta}$, where H and θ are indicated in the figure for $\omega = 300$. Since $\bar{V}_1 = 1 + j0$ for the unit voltage,

$$\bar{V}_2 = He^{-j\theta}$$

In other words, Fig. 5-23a portrays the response of the circuit for waves of di ferent frequencies and unit rms value. At several frequencies the outputs are

$$v_2 = .93\sqrt{2}\cos(300t-11^\circ) \quad \text{at} \quad \omega = 300$$

$$v_2 = .87\sqrt{2}\cos(800t-72^\circ) \quad \text{at} \quad \omega = 800$$

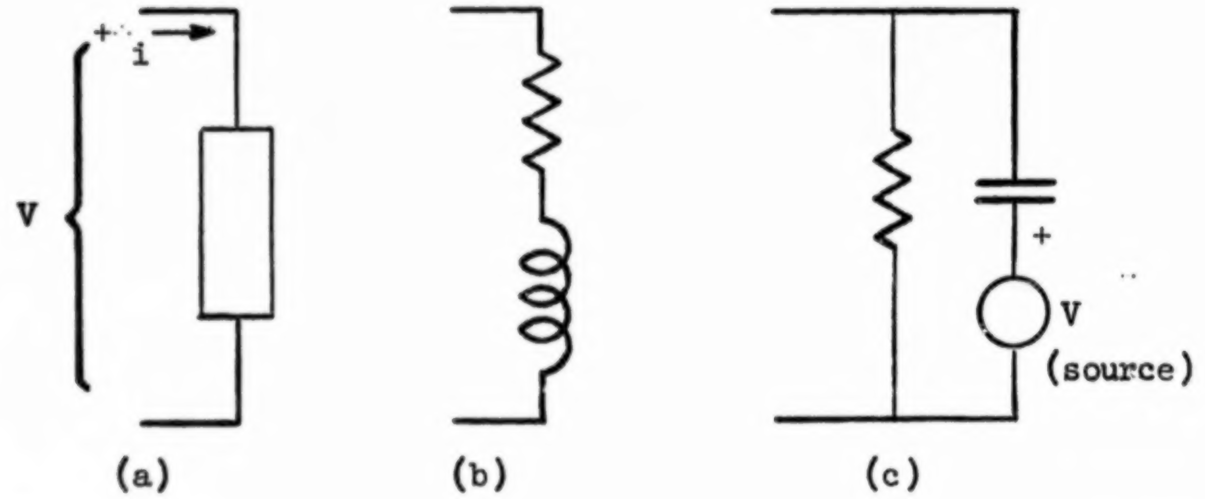
$$v_2 = .20\sqrt{2}\cos(\quad -233^\circ) \quad \text{at} \quad \omega = 2000$$

It is evident that the output decreases at the frequency increases. This why this circuit is called a low pass filter.

A common method of portraying the variation of output with frequency is tc plot H as a function of frequency, as in Fig. 5-23b. Note that logarithmic sca have been used for H and ω. This is usually done, because of the wide ranges c which they vary.

5-11 Power

Referring to Fig. 5-24a, let the box represent any combination of circuit



(a) (b) (c)

Figure 5-24.

elements, like the examples shown at (b) and (c). In general, there will be a phase difference between v and i, except in the case of a pure resistance. Thus, for the general case, v and i are respectively represented by

$$v = \sqrt{2}\, V \cos(\omega t+\beta)$$

$$i = \sqrt{2}\, I \cos(\omega t+\alpha) \qquad (5\text{-}71)$$

where the phase difference (the angle by which v leads i) is $(\beta-\alpha)$. In complex number representation, as introduced in Eqs. (5-17), these are

$$v = \sqrt{2}\left(\frac{\overline{V}e^{j\omega t} + \overline{V}^{*}e^{-j\omega t}}{2}\right)$$

$$i = \sqrt{2}\left(\frac{\overline{I}e^{j\omega t} + \overline{I}^{*}e^{-j\omega t}}{2}\right)$$

where

$$\overline{V} = Ve^{j\beta} \qquad \text{and} \qquad \overline{I} = Ie^{j\alpha}$$

To find the power, we take the product of the above and get

$$p = \frac{1}{2}\,(Ve^{j\omega t} + \overline{V}^{*}e^{-j\omega t})(Ie^{j\omega t} + \overline{I}^{*}e^{-j\omega t})$$

$$= \left(\frac{\overline{V}\overline{I}^{*} + \overline{V}^{*}\overline{I}}{2}\right) + \left(\frac{\overline{V}\overline{I}e^{j2\omega t} + \overline{V}^{*}\overline{I}^{*}e^{-j2\omega t}}{2}\right)$$

Referring to the first term on the right of the above equation, observe that

$$\overline{V}\overline{I}^{*} = VIe^{j(\beta-\alpha)} \qquad \text{and} \qquad \overline{V}^{*}\overline{I} = VIe^{-j(\beta-\alpha)}$$

so that for this term we have

$$\frac{\overline{V}\overline{I}^{*} + \overline{V}^{*}\overline{I}}{2} = VI\left(\frac{e^{j(\beta-\alpha)} + e^{-j(\beta-\alpha)}}{2}\right) = VI\cos(\beta-\alpha) \qquad (5\text{-}72)$$

In the second term observe that $\overline{V}\overline{I} = VIe^{j(\beta+\alpha)}$ and that $\overline{V}^{*}\overline{I}^{*} = (\overline{V}\overline{I})^{*} = VIe^{-j(\beta+\alpha)}$, so that for this term we have

$$\frac{\overline{V}\overline{I}e^{j2\omega t} + \overline{V}^{*}\overline{I}^{*}e^{-j2\omega t}}{2} = VI\cos(2\omega t + \beta + \alpha)$$

Thus, starting with the product vi, we have the formula for instantaneous power

$$p = VI\cos(\beta-\alpha) + VI\cos(2\omega t + \beta + \alpha) \qquad (5\text{-}73)$$

A graph of this is shown in Fig. 5-25, where the dashed line represents the constant term $P = VI\cos(\beta-\alpha)$. The $VI\cos(2\omega t+B+\alpha)$ term represents a sinusoidal wave of angular frequency 2ω, which therefore varies symmetrically with respect to this dashed line. Accordingly, over a long interval of time areas of the p curve above and below the dashed line nearly cancel, so that the dashed line is the steady state average value.[+] (The words steady state mean "average taken over a long time interval.") If P is this average, we have the result

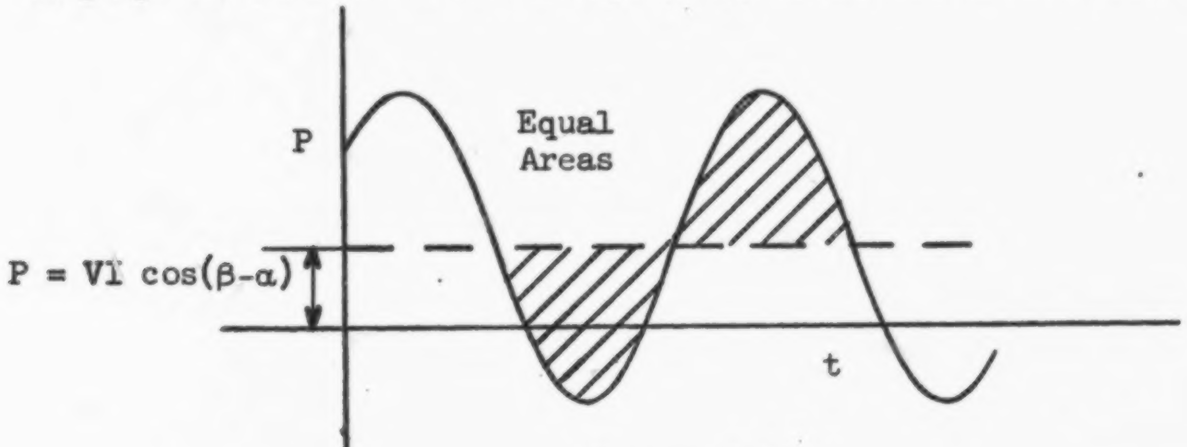


Figure 5-25.

$$P = V\ I\cos(\beta-\alpha) \qquad (5\text{-}74)$$

Observe from this that the power depends not only on the magnitude of the v and i waves, but also on the phase difference. In particular, if $\beta-\alpha = 90^\circ$, average power is zero, even if the voltage and current are not zero. This will be true of circuits having no resistance, since with only L and C elements there is a 90° phase difference between voltage and current waves.

Equation (5-72) shows that average power P can be obtained directly from $\overline{V}$ and $\overline{I}$, using

---

[+] Negative and positive areas cancel exactly only if reckoned over an integral number of cycles. If a fraction of a cycle is included in the interval over which the average is computed, the average will be different from $VI\cos\theta$. However, as the number of full cycles becomes large, the discrepancy becomes negligible.

$$P = \frac{\overline{VI}^* + \overline{V}^*\overline{I}}{2} \qquad (5\text{-}75)$$

This form is useful when $\overline{V}$ and $\overline{I}$ are obtained from previous calculations in rectangular form. For example, suppose

$$\overline{V} = V_1 + jV_2 \qquad \text{and} \qquad \overline{I} = I_1 + jI_2$$

Then,

$$\overline{VI}^* = (V_1+jV_2)(\overline{I}_1-j\overline{I}_2) = V_1I_1 + V_2I_2 + j(V_2I_1-V_1I_2)$$

$$\overline{V}^*\overline{I} = (V_1-jV_2)(I_1+jI_2) = V_1I_1 + V_2I_2 - j(V_2I_1-V_1I_2)$$

and thus, by substituting the sum of the above in Eq. (5-70), we have

$$P = V_1I_1 + V_2I_2 \qquad (5\text{-}76)$$

This last formula is easily remembered. If voltage and current phasors are known in rectangular form, (i.e., real and imaginary parts are explicitly known) the average power is obtained by multiplying the real parts of $\overline{V}$ and $\overline{I}$, multiplying the imaginary parts, and adding the two quantities so obtained. As an alternate viewpoint, observe that the right-hand side of Eq. (5-76) is identical with the real part of either $\overline{VI}^*$ or $\overline{V}^*\overline{I}$. Accordingly, Eq. (5-75) can be replaced by

$$P = \mathrm{Re}(\overline{VI}^*)$$

where the symbol Re( ) means the real part of the quantity in the parentheses.

The product VI (which would be power in a d-c circuit) is called apparent power in steady state circuit analysis. The ratio

$$F = \frac{P}{VI} = \frac{\text{average power}}{\text{apparent power}} \qquad (5\text{-}77)$$

is called the power factor of the device or circuit branch in question. From Eq. (5-74) it is evident that

$$F = \cos(\beta-\alpha)$$

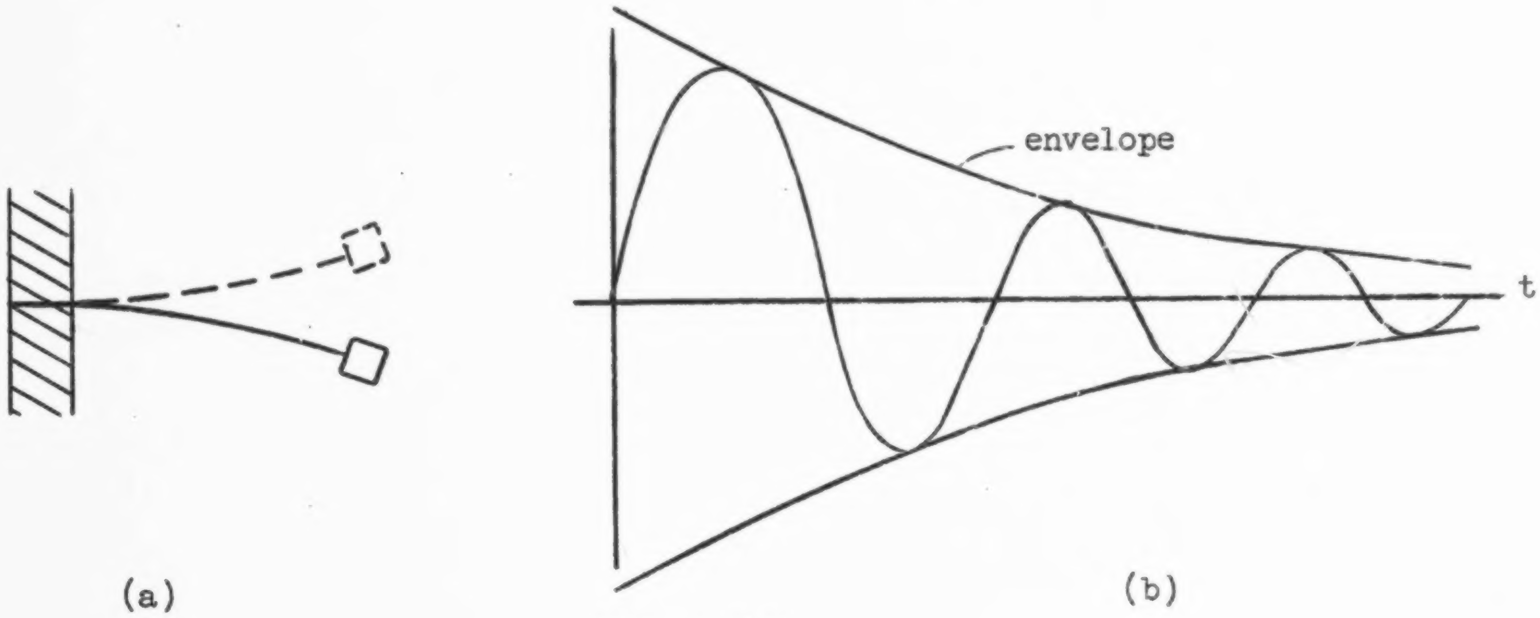
where $(B-\alpha)$ is the phase difference between the v and i waves. Apparent power is measured in volt amperes.

Chapter 6

NATURAL RESPONSE OF ELECTRIC CIRCUITS

Introduction

In the everyday world there are many opportunities to observe vibrations in mechanical systems. For example, a weight at the free end of a strip of springy steel clamped at the other end, as in Fig. 6-1a, will vibrate if it is deflected and then released. A plot of the deflection as a function of time is shown at (b). This is called a damped oscillation or a damped sinusoid, because the oscillations gradually die out. You will observe that the oscillation occurs in the absence of any kind of oscillatory source. The spring is merely deflected, and then let go. It is, therefore, appropriate that the resulting wave should also be called a natural response or a natural oscillation. There are many other examples: the oscillation of a piano string, the oscillation of the pointer on a weighing scale when a weight is suddenly placed on it, etc. In the case of certain electrical circuits, currents, voltages, and charges can also oscillate in a similar manner.



(a) (b)

Figure 6-1.

1 A constant amplitude sinusoidal wave has the equation

$$v = V_m \cos(\omega t+\beta)$$

as we saw in Chap. 5. A damped sinusoid has the equation

2 $$v = V_m e^{-at} \cos(\omega t+\beta)$$

That is, it is a sinusoidal wave multiplied by a factor e^{-at} which forces the waves to be "pinched" down between the envelope curves shown in Fig. 6-1b. As time goes on, the amplitude of the wave decreases or decays.

3 In this chapter we shall show how, under certain conditions, responses of the above form are obtained from electric circuits. We shall also find that under certain conditions, if the rate of decay exceeds a certain value, the oscillatory character of the response disappears, and the above response
4 function reduces to a sum of two exponentials, like

$$v = V_a e^{-at} + V_b e^{-bt}$$

where V_a, V_b, a, and b are constants.

In the body of this chapter we shall derive these equations, determine
5 how values of the circuit constant determine whether or not the response is oscillatory, and show how to obtain the various constants appearing in these equations.

6-1 The Circuit Equations

6 A number of different networks could be used as an introductory illustration. The series circuit shown in Fig. 6-2 is chosen. After the switch is closed, Kvl applies, and we have

$$v_L + v_R + v_C = V_B$$

7 and thus

$$L \frac{di}{dt} + Ri + v_C = V_B \qquad (6\text{-}1)$$

In addition to this equation, i and v_C are related by

$$C \frac{dv_C}{dt} = i \qquad (6\text{-}2)$$

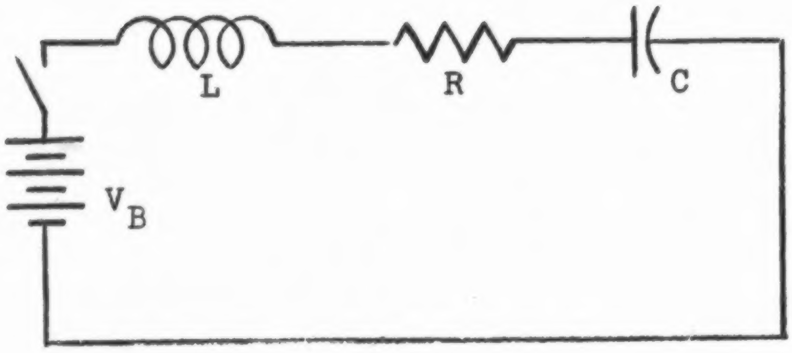


Figure 6-2.

Equations (6-1) and (6-2) constitute a pair of simultaneous equations in the two variables v_C and i. However, they are differential equations rather than algebraic equations. It is our objective to find solutions for these equations.

Ensuing steps will be simplified by introducing two new variables

$$x_1 = \sqrt{\frac{L}{C}}\, i \qquad \text{(a)}$$

$$(6\text{-}3)$$

$$x_2 = v_C - V_B \qquad \text{(b)}$$

and the parameters

$$\alpha = \frac{R}{L} \qquad \text{(a)}$$

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad \text{(c)}$$

1 Incorporating these notational changes in Eqs. (6-1) and (6-2) yields the two new equations

$$\frac{dx_1}{dt} = -\alpha x_1 - \omega_0 x_2 \tag{6-5}$$

2

$$\frac{dx_2}{dt} = \omega_0 x_1 \tag{6-6}$$

as you can easily show by carrying out the steps yourself. These variable changes are not necessary to obtain a solution, but they have the advantage of reducing the original equations to the somewhat simpler forms shown above.
3 The remainder of our discussion will apply to these expressions, with the understanding that Eqs. (6-3) can be used to regain the original quantities v_C and i. You should observe that both x_1 and x_2 have the dimensions of voltage.

4 A two-pronged attack will be used. The first will be a graphical analysis of the equations, from which physical insight can be gained as to the nature of the solutions. The second will be the actual mathematical solution. While both are important, the first part of the discussion is
5 particularly significant, from an engineering viewpoint.

6-3 The Phase Plane

Since we are dealing with the two variables x_1 and x_2, it is natural to think of plotting their relationship in a plane, called a phase plane. These quantities vary with time, and at any instant of time the pair of
6 numbers (x_1, x_2) completely specifies the state of the circuit. Consequently, these variables are called the state variables. The point in the plane having the coordinates (x_1, x_2) is called the state point. In thinking
7 about the phase plane we use the mental image of an ordinary geometrical plane surface. Thus, we talk about distances in the plane, even though the dimensions of the axis are not length, and we talk about velocities as the phase point moves around in the plane.

Let us determine how the state point varies with time, for the simplest possible case ($R = 0$, which means $\alpha = 0$), so that the equations reduce to

$$\frac{dx_1}{dt} = - \omega_0 x_2$$
$$\frac{dx_2}{dt} = \omega_0 x_1 \tag{6-7}$$

To begin, we observe that an explicit algebraic solution is impossible, because these are differential equations. However, assume the state of the system is known at some particular instant of time; that is, assume x_1 and x_2 are known. Algebra can then be used to determine how the state is changing at that instant, by finding the velocity of the state point. This can be done because dx_1/dt and dx_2/dt are, respectively, the x_1 and x_2 components of that velocity, as shown by the two dashed vectors in Fig. 6-3.

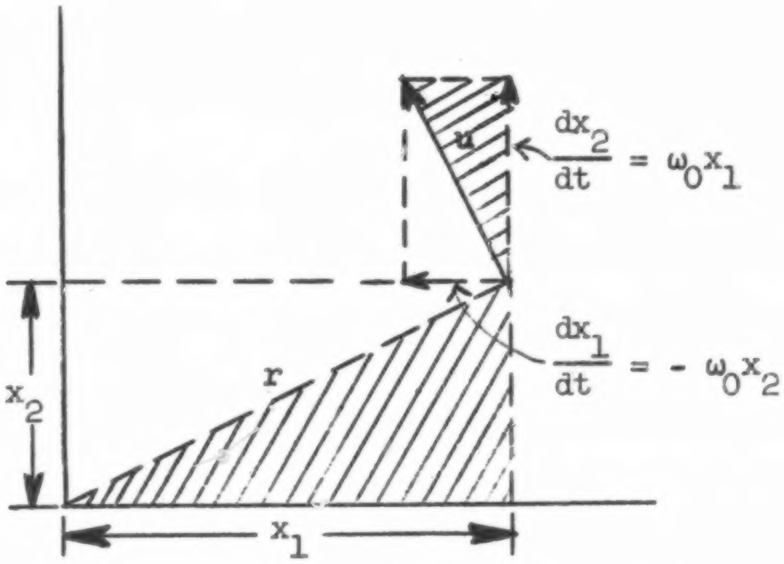

u
$\frac{dx_2}{dt} = \omega_0 x_1$
r
x_2
$\frac{dx_1}{dt} = - \omega_0 x_2$
x_1

Figure 6-3.

The actual velocity of the point is indicated by the solid arrow, and this velocity has a magnitude which can be found from the figure to be

$$u = \sqrt{\left(\frac{dx_1}{dt}\right)^2 + \left(\frac{dx_2}{dt}\right)^2} = \omega_0 \sqrt{x_1^2 + x_2^2}$$

1 From Fig. 6-3 we make the following observations:

(1) $\sqrt{x_1^2 + x_2^2} = r$, the distance from the origin to the point;

(2) The shaded triangles are similar, and therefore the velocity is perpendicular to the radial line r.

2 Now observe that ω_0 ($= 1/\sqrt{LC}$) is a constant, and that

$$u = r\omega_0 \tag{6-8}$$

Since the point moves with a velocity perpendicular to a radius, it follows that the point moves on a circle, with radius r. (Why?) Thus, we find that
3 the velocity of the state point has a constant magnitude. The point seems to be rigidly attached to a wheel rotating with constant angular velocity ω_0.

The problem is not completely solved, but we have taken a big step. Although ω_0 is known from the circuit parameters, we do not yet know a value of r, nor is it clear where the point is at various instants of time.

4 Completion of the solution requires knowledge about the circuit at some particular instant of time, usually taken as t = 0. Thus, let $x_1(0)$ and $x_2(0)$ be the values of the state variables when t = 0. These values, which are called the initial conditions, must be determined from the specified problem. For example, suppose the actual circuit arrangement is shown in Fig. 6-4, in
5 which the switching arrangement is slightly more complicated than in Fig. 6-2, in order to provide the possibility of having an initial current i(0) in the conductor. The capacitor is assumed to have a charge at t = 0, leading to an initial capacitor voltage $v_C(0)$. Thus, the initial values of the two state variables are

6
$$x_1(0) = \sqrt{\frac{L}{C}}\, i(0) \qquad \text{(a)}$$

$$\text{(6-9)}$$

$$x_2(0) = v_C(0) - V_B \qquad \text{(b)}$$

7 As battery A is disconnected from the inductance, at t = 0, the inductance cannot have a discontinuity of current, and so i(0) must become the current in the L-C circuit. Although this may seem to be a rather artificial situation, it provides the possibility that $x_1(0)$ and $x_2(0)$ can both be different from zero.
8 We could have simply said assume both $x_1(0)$ and $x_2(0)$ are non-zero.

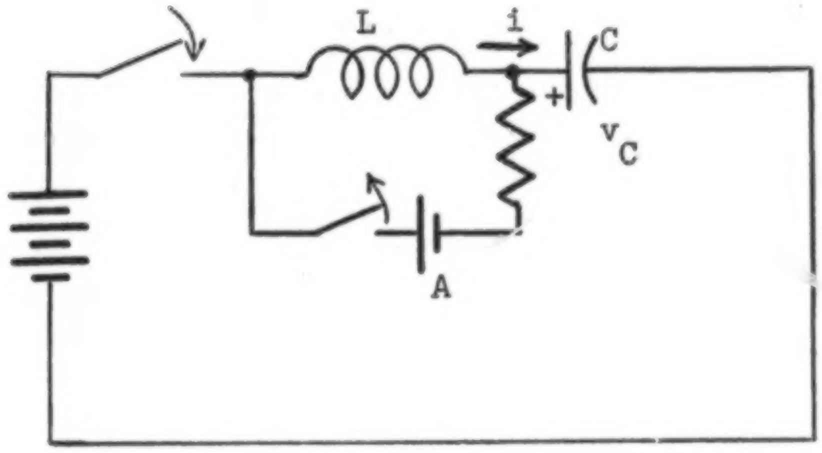


Figure 6-4.

With knowledge of $x_1(0)$, and $x_2(0)$, the position of the rotating point is known at $t = 0$, as indicated in Fig. 6-5, in which the state point is the tip of the radius line. Since we also know that the line rotates with angular velocity ω_0, the motion of the state point becomes completely known. The time variation of $x_1(t)$ and $x_2(t)$ can now be obtained graphically, as shown in Fig. 6-5. This construction has an unusual orientation, with the time axis in a vertical position for x_2.

Analytical expressions are easily obtained from this graphical analysis. The radius

$$r = \sqrt{[x_1(0)]^2 + [x_2(0)]^2}$$

and angle

$$\theta = \arctan \frac{x_2(0)}{x_1(0)} \qquad (6\text{-}10)$$

are known from the initial conditions. Also, at any time t the angular position of the state point is $\omega t + \theta$, and so we have

$$x_1(t) = \sqrt{[x_1(0)]^2 + [x_2(0)]^2} \cos \left[\omega_0 t + \arctan \frac{x_2(0)}{x_1(0)}\right] \qquad \text{(a)}$$

$$x_2(t) = \sqrt{[x_1(0)]^2 + [x_2(0)]^2} \sin \left[\omega_0 t + \arctan \frac{x_2(0)}{x_1(0)}\right] \qquad \text{(b)} \qquad (6\text{-}11)$$

1 It would have been rather easy to obtain these equations mathematically
from the original equations. However, this graphical approach has the significant
advantage of providing a pictorial indication of how the response is determined
by the initial conditions, through their effect on both the radius of the
2 rotating line and its initial angle.

This example has pedagogical value because, as the limiting case ($\alpha = 0$)
in which there is no energy dissipation, it provides a base for consideration
of the more general case. The $\alpha = 0$ case can also be of direct practical
3 use in the analysis of transistor or vacuum tube oscillators. We shall not go
into the details of this question here, except to point out that an essential
function of the vacuum tube or transistor in an oscillator is to cancel the
effect of the actual circuit resistance.

4 The fact that $x_1^2 + x_2^2$ is a constant has interesting implications in
the case where $V_B = 0$, so that $x_2 = v_C$. Going back to the definitions of
x_1, and x_2, in Eqs. (6-3), we then have

5 $$\frac{L}{C} i^2 + v_C^2 = r^2$$

or

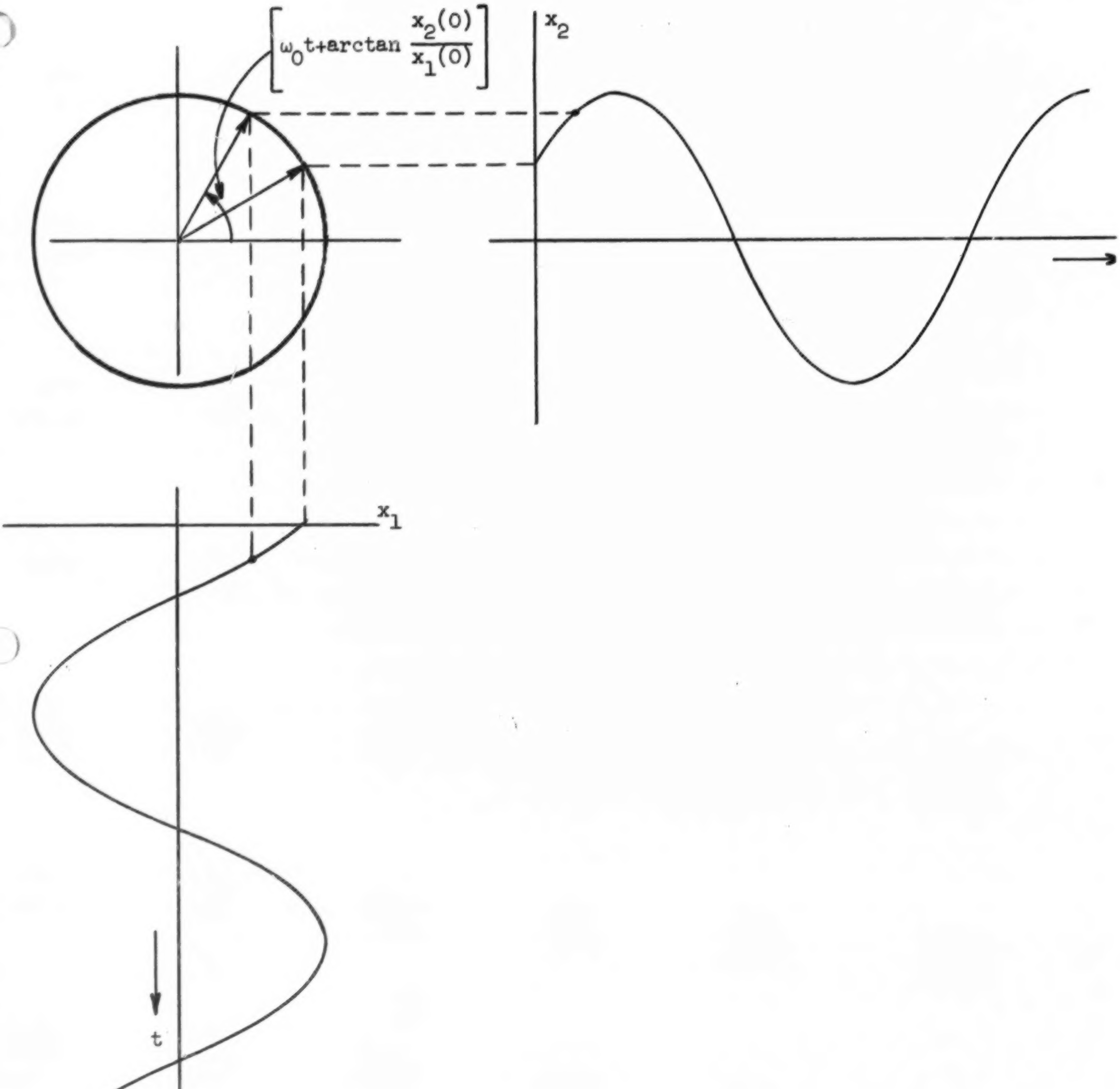
$$\frac{Li^2}{2} + \frac{Cv_C^2}{2} = \frac{Cr^2}{2} \qquad (6\text{-}12)$$

The right-hand side of this equation is a constant. On the left, $Li^2/2$ is
6 the energy stored in the inductanct at any instant, and $Cv_C^2/2$ is
the energy stored in the capacitance. The total energy is constant (as we
should expect, since no energy is dissipated in the absence of resistance),
and so as one increases the other decreases. Twice during each cycle the
7 energy is completely interchanged between the inductance and capacitance,
being a maximum in one when it is zero in the other.

Now suppose $\alpha \neq 0$. Continuing to think of the velocity of the state
point in the phase plane, referring to Eqs. (6-5) and (6-6) we see that
8 the vertical velocity dx_2/dt is unchanged, but that the horizontal component
now becomes

$$\frac{dx_1}{dt} = -\alpha x_1 - \omega_0 x_2$$

In other words, at any point not on the x_2 axis an additional component
9 $-\alpha x_1$ must be added to the horizontal velocity, as indicated by two examples


$\left[\omega_0 t + \arctan \frac{x_2(0)}{x_1(0)}\right]$
x_2
x_1
t

in Fig. 6-6, one at a positive value of x_1 and one at a negative value. The additional velocity component is always toward the x_2 axis, and is proportional to x_1, the distance of the point from the x_2 axis. The - sign on $-\alpha x_1$ ensures that the direction of this additional velocity component shall be toward the x_2 axis, regardless of the location of the state point.

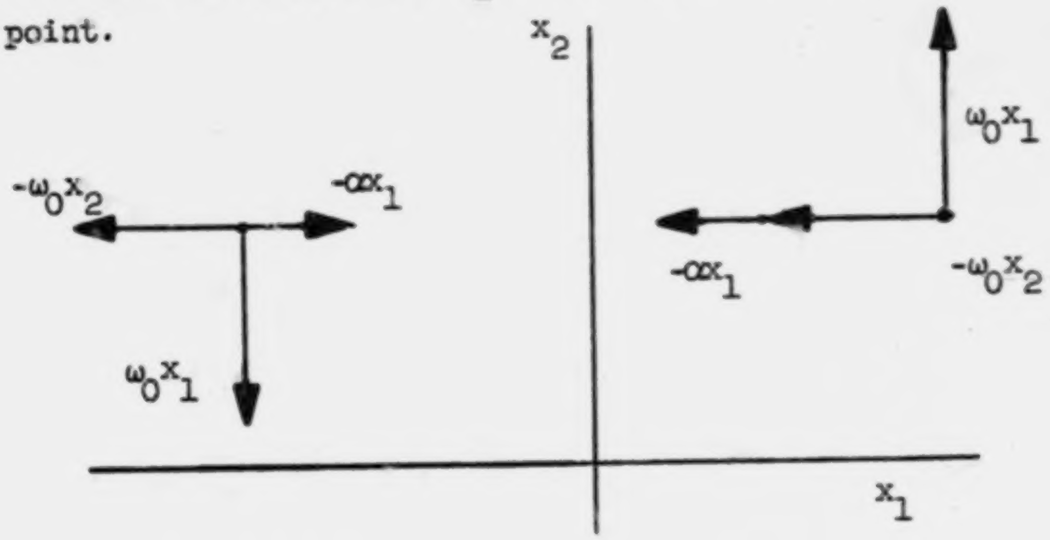


Figure 6-6.

For analysis of this case, it is convenient to write the original equations in the form

$$\left(\frac{1}{\omega_0}\right)\frac{dx_1}{dt} = -\left(\frac{\alpha}{\omega_0}x_1 + x_2\right) \qquad (6\text{-}13)$$

$$\left(\frac{1}{\omega_0}\right)\frac{dx_2}{dt} = x_1$$

This shows that the ratio [illegible]

$$\frac{\alpha}{\omega_0} = \frac{R}{L}\sqrt{LC} = R\sqrt{\frac{C}{L}}$$

determines the extent to which the velocity is modified. A good idea of how the circuit will respond can be obtained by calculating dx_1/dt and dx_2/dt at a variety of points, and plotting the resulting velocity vectors. Figures 6-7 and 6-8 show these sets of vectors for the two cases $\alpha = \omega_0/2$ and $\alpha = 2\omega_0$.

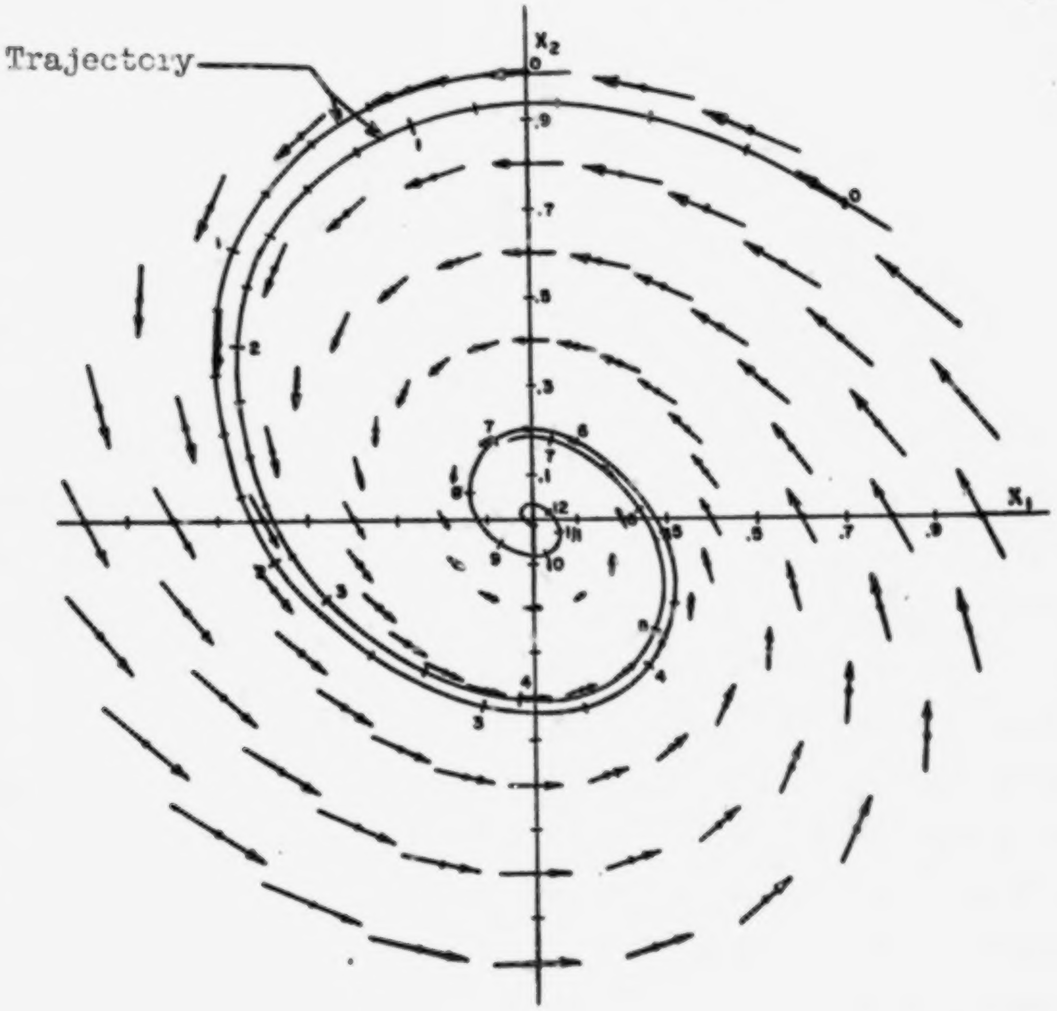

Trajectory
x_2
x_1
0
.9
.7
.5
.3
.1
.1
.3
.5
.7
.9
0
1
2
3
4
5
6
7
8
9
10
11
12

Fig. 6-7

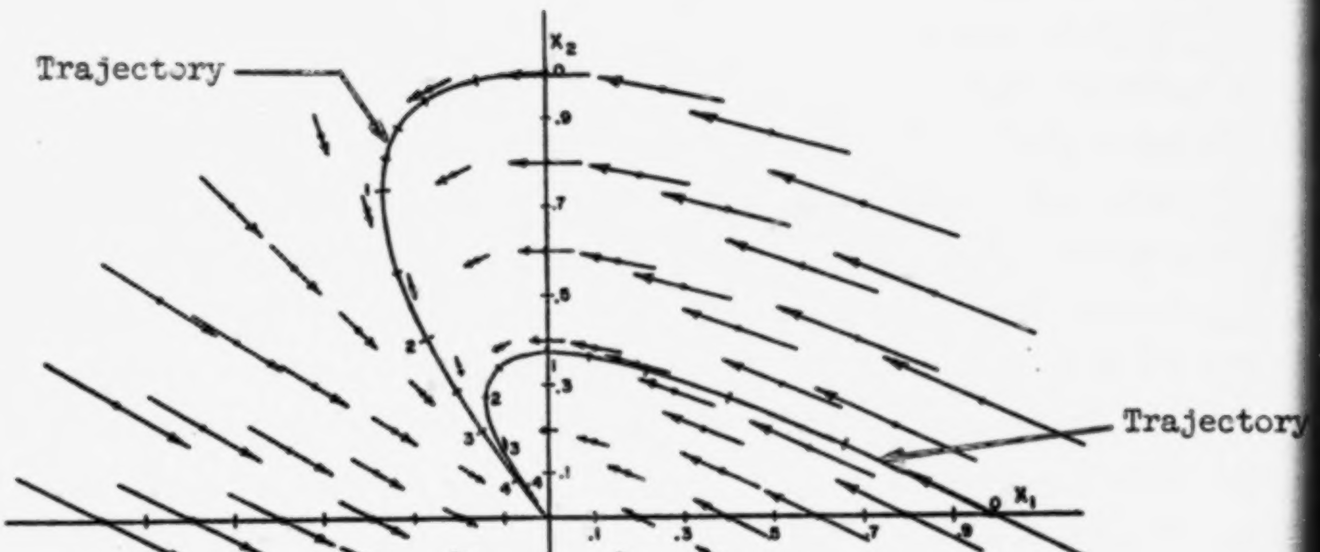

Trajectory
Trajectory
x_2
x_1
0
.9
.7
.5
.3
.1
.1
.3
.5
.7
.9
0
1
2
3
4

Fig. 6-8

Such plots as these do not provide explicit solutions, but they do provide considerable insight as to what happens. As in the case when $\alpha = 0$, initial values $x_1(0)$ and $x_2(0)$ determine a starting point, from which the point moves with the appropriate velocity for that position. As the point moves, the velocity also changes and there is a new velocity appropriate for each position, and a path, or trajectory, is traced out. Because the velocities are tilted toward the vertical axis, these trajectories generally go toward the origin. Physically, this means that the circuit approaches a state of rest ($x_1 = x_2 = 0$, meaning $i = 0$ and $v_C = V_B$) as energy is gradually dissipated in the resistor. Several such trajectories are shown in Figs. 6-7 and 6-8 for different initial conditions.

These trajectories do not show the time variation directly, but this can be included (preferably in terms of values of $\omega_0 t$) by placing marks on the trajectory. In these figures, each arrow is of a lenght 0.2u, meaning its length is the distance the point will move from the dot at the center of the arrow when $\omega_0 t$ increases by an amount 0.2, assuming a uniform velocity for this short length of time. Such time interval markings are included on the trajectories. Graphs of x_1 and x_2 as functions of time can be obtained by projecting points, as indicated in Figs. 6-9 and 6-10.

In comparing the cases $\alpha = \omega_0/2$ and $\alpha = 2\omega_0$, it is seen that in the former case the responses are basically oscillatory, but that this is not true when $\alpha = 2\omega_0$. The energy interpretation given for the dissipationless case provides some insight as to the reason for this difference. As previously stated, when there is no dissipation, energy is passed back and forth between capacitance and inductance, while the total energy remains constant. When there is dissipation ($\alpha > 0$), some energy is lost during each interchange, and hence each successive voltage and current peak of Fig. 6-9 is smaller than the previous one. However, when α becomes equal to or larger than a certain value (called the critical value, which we shall presently show is $2\omega_0$), the energy dissipation becomes so large that a complete cycle of operation never occurs, and the result is a response, as in Fig. 6-10, which is essentially non-oscillatory.

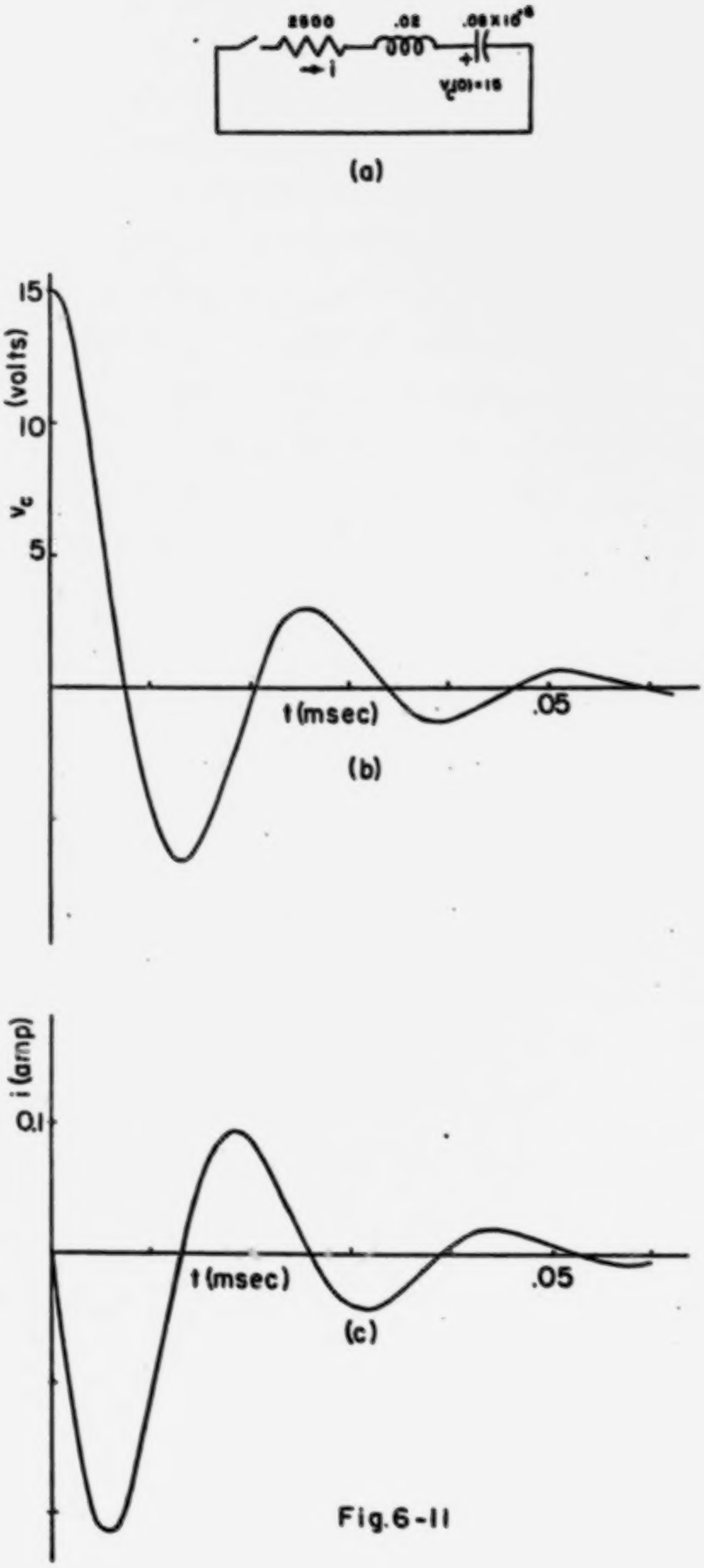


Fig. 6-11

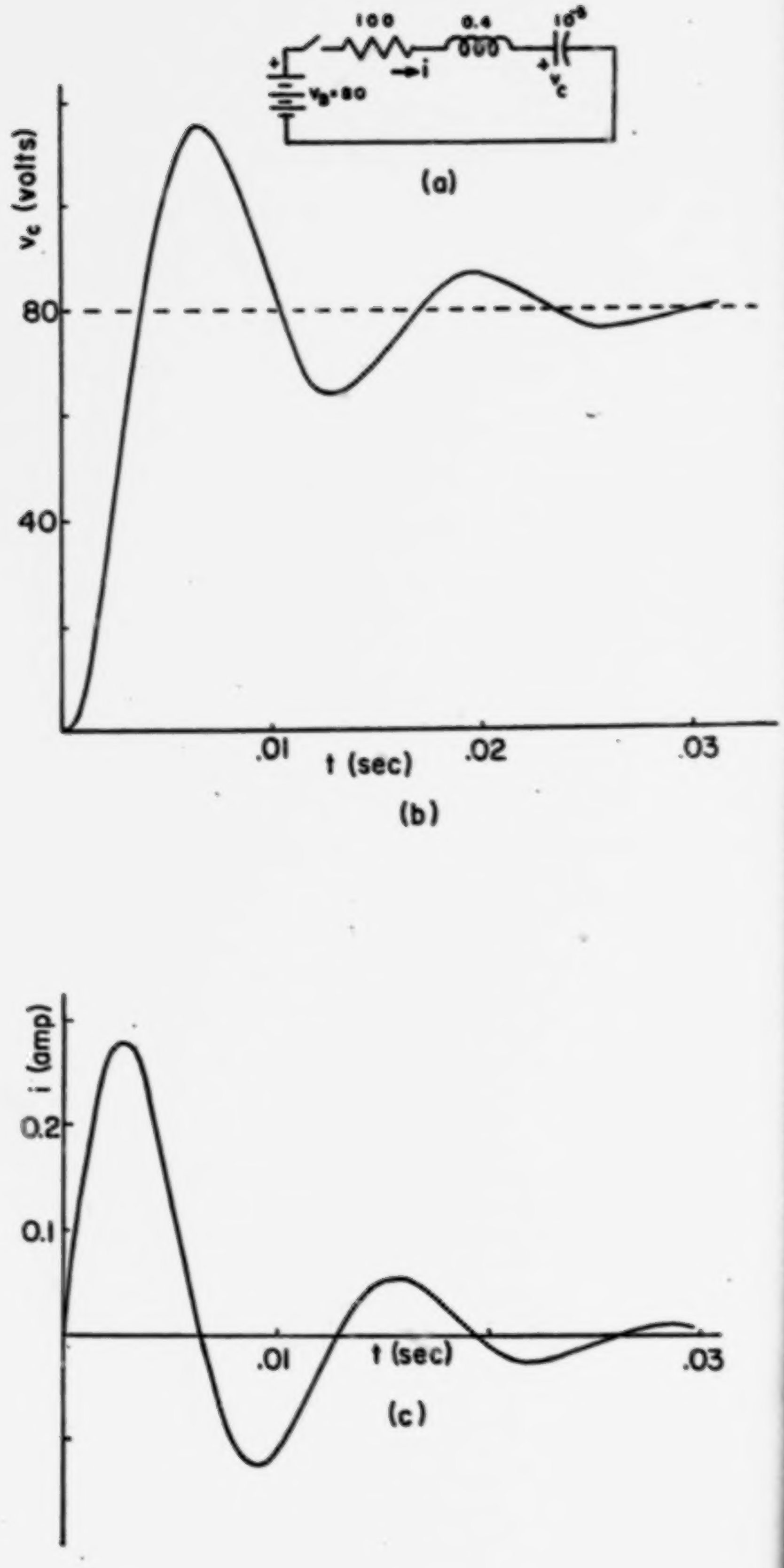


Fig. 6-12

This phase plane interpretation does not provide exact solutions, but it does show the effect of circuit resistance, and, as in the case of Fig. 6-5, shows how initial conditions determine the response curve, merely by locating the state point at t = 0. When R, L, and C are constants, as we have tacitly assumed, exact analytical solutions are also possible, as we shall presently show. However, the phase plane approach can also be used for nonlinear equations, where α and/or ω_0 are functions of x_1 and/or x_2. Such a functional relationship merely modifies the velocity vector at each point in the phase plane. Graphical analyses are then generally necessary, or approximate point-by-point numerical calculations can be made to determine a succession of state points, perhaps by digital computer methods. The extensive velocity plots of Figs. 6-7 and 6-8 are not needed to obtain solutions by this method, but they are included here for their pedagogical value, to aid in your visualization of a variety of trajectories from various starting points.

Illustrative Example

For the trajectories given in Figs. 6-7 and 6-8, the maximum radial coordinate for each curve (which occurs at the initial point) is unity. For this reason, they can be called normalized curves. Obviously, practical problems can have a value (r_m) for this initial radius which is different from unity. Curves like Figs. 6-7 and 6-8 can then be regarded as plots of x_1/r_m vs. x_2/r_m (which will have a maximum value of unity). Actual values of x_1 and x_2 can then be obtained from the normalized curve merely by multiplying by r_m.

As an example, consider the circuit of Fig. 6-11a. It is assumed that the capacitor is initially charged to 15 volts, and we are to determine the variation of v_C and i after the switch is closed. Using the numerical values shown on the figure, we have

$$\alpha = \frac{2500}{.02} = 1.25 \times 10^5$$

$$\omega_0 = \frac{1}{\sqrt{(.02)(.08)\times 10^{-6}}} = 2.5 \times 10^5$$

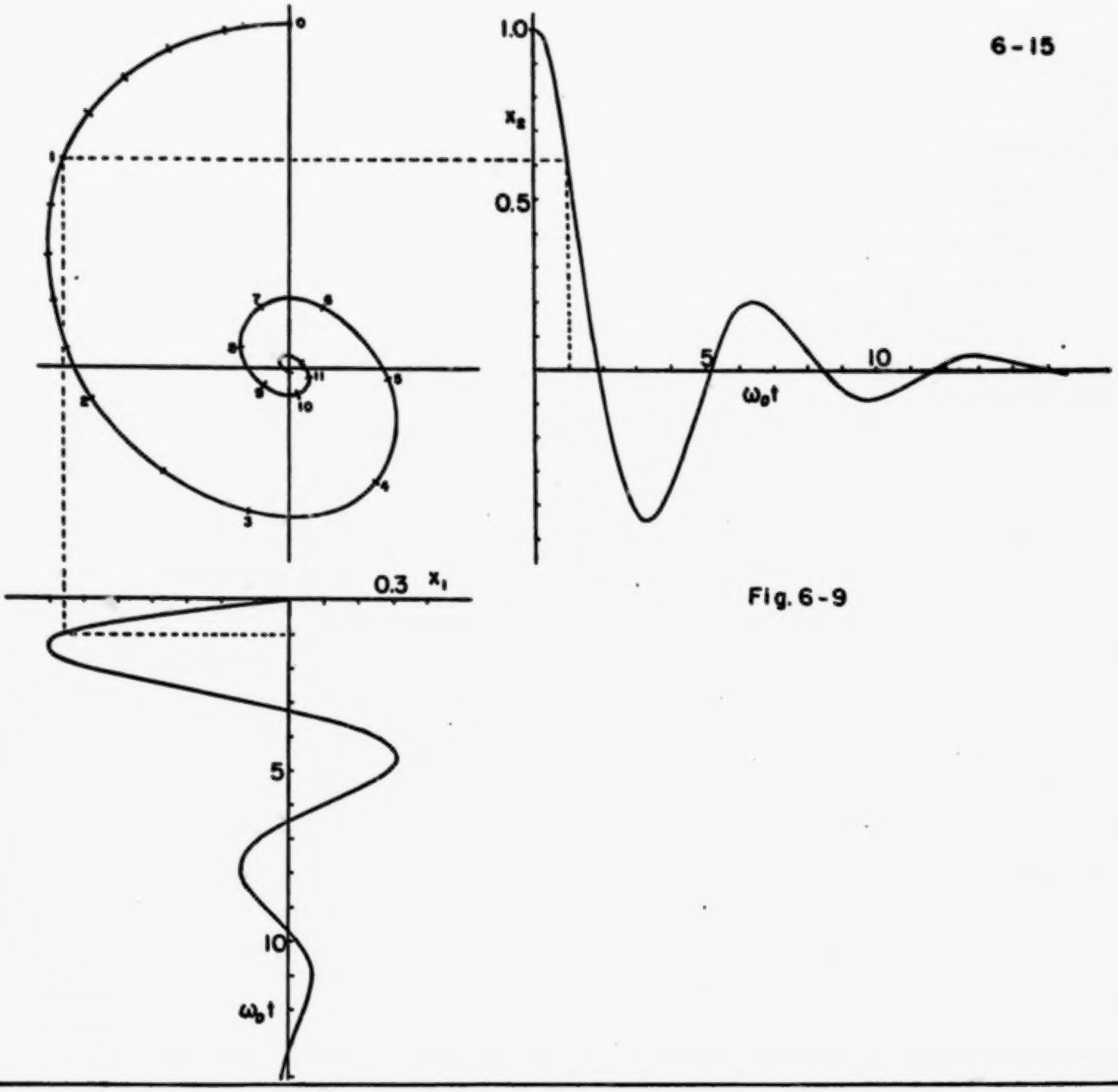


Fig. 6-9

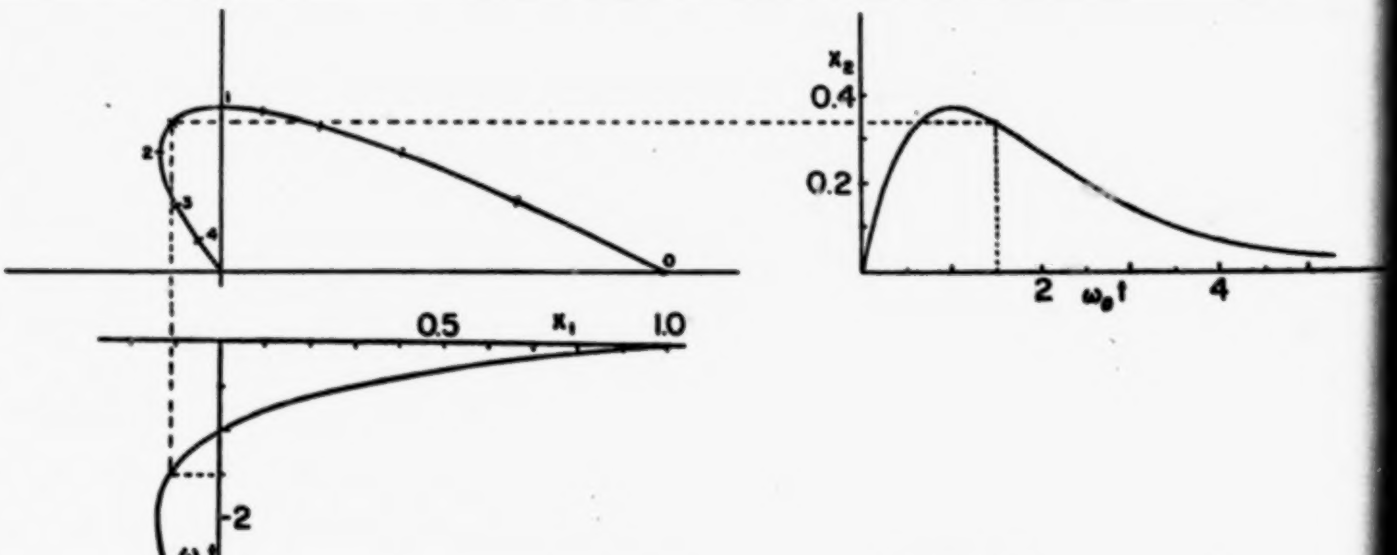


Fig. 6-10

Thus, we see that $\alpha = \omega_0/2$, and therefore that Fig. 6-7 is applicable. Also,

$$x_1 = \sqrt{\frac{.02}{.08\times 10^{-6}}}\; i = 500i$$

$$x_2 = v_C$$

and, from the stated initial conditions, $r_m = \sqrt{x_1(0)^2 + x_2(0)^2} = 15$. Since $x_1(0) = 0$, we see that Fig. 6-9, which was derived from Fig. 6-7, is the applicable curve, if values of x_2 are multiplied by 15. Values of x_1 are also multiplied by 15, but if we wish to obtain current i, we also divide by 500. Thus, to obtain i, the x_1 ordinates labels of Fig. 6-9 are multiplied by 15/500 = .03. As mentioned previously, the time scale is obtained by dividing values of $\omega_0 t$ by $\omega_0 = 2.5 \times 10^5$ in this example. The resulting curves are shown in Fig. 6-11 (b) and (c). These are the same as the curves in Fig. 6-9, except for the scale changes noted above.

Another example is shown in Fig. 6-12a. In this case the capacitor is initially uncharged, and a battery of voltage V_E = 80 volts is switched into the circuit at t = 0. For this example

$$\alpha = \frac{100}{.4} = 250$$

$$\omega_0 = \frac{1}{\sqrt{(.40)(10)\times 10^{-6}}} = 500$$

Again $\alpha = \omega_0/2$, and so Fig. 6-7 can be used. The variables are

$$x_1 = \sqrt{\frac{.4}{10\times 10^{-6}}}\; i = 200i$$

$$x_2 = v_C - 80$$

Observe that $x_1(0) = 0$ and $x_2(0) = -80$, and $r_m = 80$. We do not have a curve starting at a negative value of x_2. However, Fig. 6-7 is symmetrical, and so the trajectory starting at $x_2(0) = 1$ can be used if signs on x_1 and x_2 are reversed. Furthermore, these variables are to be increased by the factor 80.

Thus, in effect x_1 and x_2 in Figs. 6-7 and 6-9 should be multiplied by -80 to yield the above values of x_1 and x_2. Furthermore, to get i from the coordinates of Fig. 6-9, we divide x_1 by 200, and to get v_C we add 80. By inverting the x_2 of Fig. 6-9 and adding 80, we obtain a curve for v_C which starts from the origin and gradually settles down to the constant value 80, as shown in Fig. 6-12b. In other words, from the physical viewpoint, the capacitor is eventually charged to voltage $V_B = 80$. The current curve is shown in Fig. 6-12c, and this offers an interesting comparison with Fig. 6-11c. In Fig. 6-11 the current starts in a negative direction, because the capacitor is initially charged positively. When the switch is closed, positive charge begins to leave the left-hand plate, causing i to be negative. In Fig. 6-12, the capacitor voltage starts from zero, and so the applied voltage of the battery causes current to flow in the direction of the arrow, making i positive.

6-4 Mathematical Solution

We shall now consider the exact solution of the original equations, namely

$$\frac{dx_1}{dt} = -\alpha x_1 - \omega_0 x_2$$

$$\frac{dx_2}{dt} = \omega_0 x_1$$

for the linear case (α and ω_0 are independent of x_1 and x_2).

The form of the solutions $x_1(t)$ and $x_2(t)$ of these equations is not evident. However, it is a legitimate procedure to solve equations by a trial method, checking the trial solutions to see if they are satisfactory. The trial solutions will be

$$x_1 = A_1 e^{st}$$
$$x_2 = A_2 e^{st} \tag{6-14}$$

where s is an undetermined constant introduced to provide generality to the exponential function. (That is, e^{st} is much like the exponentials $e^{-t/T}$ encountered in R-C and R-L networks, s being similar to $-1/T$.) These are readily differentiated and substituted into the original equations, with the result

$$sA_1e^{st} = -\alpha A_1e^{st} - \omega_0A_2e^{st}$$

$$sA_2e^{st} = \omega_0A_1e^{st}$$

The common factors e^{st} can be cancelled, giving

$$\begin{aligned} (s+\alpha)A_1 &= -\omega_0A_2 & \text{(a)} \\ sA_2 &= \omega_0A_1 & \text{(b)} \end{aligned} \qquad (6\text{-}15)$$

We now think of s as any constant, and attempt to solve these equations for A_1 and A_2. The second equation gives $A_2 = (\omega_0/s)A_1$ which, when substituted in the first yields

$$(s+\alpha)A_1 = -\frac{\omega_0^{\ 2}}{s}A_1 \quad \text{or} \quad (s+\alpha+\frac{\omega_0^{\ 2}}{s})A_1 = 0$$

$A_1 = 0$ is a solution, but this is trivial and of no interest. If A_1 is not zero, we see that s must have a value which is a solution of the quadratic equation

$$s^2 + \alpha s + \omega_0^{\ 2} = 0 \qquad (6\text{-}16)^*$$

This is called the characteristic equation of the network, and it has the two roots

---

[*] In more advanced treatments, where more than two equations is a possibility, Eqs. (6-15) are written

$$(s+\alpha)A_1 + \omega_0A_2 = 0$$

$$-\omega_0A_1 + sA_2 = 0$$

An attempt to solve these for A_1 and A_2 by using Cramer's rule, gives

$$A_1 = \frac{\begin{vmatrix} 0 & \omega_0 \\ 0 & s \end{vmatrix}}{\begin{vmatrix} s+\alpha & \omega_0 \\ -\omega_0 & s \end{vmatrix}}$$

Since the numerator is zero, A_1 can have a non-zero value only if

$$\begin{vmatrix} s+\alpha & \omega_0 \\ -\omega_0 & s \end{vmatrix} = 0$$

Expansion of this determinant yields Eq. (6-16).

1

$$\begin{aligned} s_a &= -\frac{\alpha}{2} + \sqrt{\frac{\alpha^2}{4} - \omega_0^2} = -\frac{\alpha}{2} + \omega_0 \sqrt{\left(\frac{\alpha}{2\omega_0}\right)^2 - 1} \\ s_b &= -\frac{\alpha}{2} - \sqrt{\frac{\alpha^2}{4} - \omega_0^2} = -\frac{\alpha}{2} - \omega_0 \sqrt{\left(\frac{\alpha}{2\omega_0}\right)^2 - 1} \end{aligned} \qquad (6\text{-}17)$$

2 which are different, unless $\alpha = 2\omega_0$. We shall exclude this exceptional case, for the present. Roots s_a and s_b are called the eigenvalues of the original equations. (The prefix eigen comes from a German word meaning characteristic.)

The two quantities s_a and s_b were obtained through our attempts to solve
3 for A_1 and A_2 from Eqs. (6-15). From the second of these equations, we find

$$A_2 = \frac{\omega_0}{s_a} A_1 \qquad \text{or} \qquad A_2 = \frac{\omega_0}{s_b} A_1$$

4 depending on which eigenvalue is used. (The other equation also leads to the same results with a certain amount of mathematical manipulation. (Can you do it?) Let s_a be chosen. Then, from Eqs. (6-14) we see that

5

$$\begin{aligned} x_1 &= A_1 e^{s_a t} \\ x_2 &= \left(\frac{\omega_0}{s_a}\right) A_1 e^{s_a t} \end{aligned} \qquad (6\text{-}18)$$

6 are solutions of the original equations. Constant A_1 is arbitrary. (A direct check is worthwhile.) Since A_1 is arbitrary, and since s_b might just as well have been used in Eqs. (6-18), it follows that

7

$$\begin{aligned} x_1 &= B_1 e^{s_b t} \\ x_2 &= \left(\frac{\omega_0}{s_b}\right) B_1 e^{s_a t} \end{aligned}$$

8 are also solutions, where B_1 is another arbitrary constant. When equations are linear, as is the case here, the sum of two solutions is also a solution. (An easy way to see this is to try it.) Thus, we can now say that the complete solutions are

9

$$\begin{aligned} x_1 &= A_1 e^{s_a t} + B_1 e^{s_b t} && \text{(a)} \\ x_2 &= \left(\frac{\omega_0}{s_a}\right) A_1 e^{s_a t} + \left(\frac{\omega_0}{s_b}\right) B_1 e^{s_b t} && \text{(b)} \end{aligned} \qquad (6\text{-}19)$$

These are solutions in the sense that identities are obtained upon substituting them into the original equations. However, the arbitrary constants A_1 and B_1 must be evaluated in order for Eqs. (6-19) to be solutions for a particular case. The fact that there are two constants in Eqs. (6-19) corresponds to the fact, as explained in the discussion of phase plane solutions, that there are two initial conditions, $x_1(0)$ and $x_2(0)$. Thus, for a specific case, placing $t = 0$ in Eqs. (6-19) gives

$$x_1(0) = A_1 + B_1 \qquad \text{(a)}$$

$$x_2(0) = \left(\frac{\omega_0}{s_a}\right)A_1 + \left(\frac{\omega_0}{s_b}\right)B_1 \qquad \text{(b)} \qquad (6\text{-}20)$$

as two equations which can be solved for A_1 and B_1.

This completes the formal steps in arriving at the solution, with the exception that the special case $s_a = s_b$ has not been treated. We can see that Eqs. (6-19) do not then provide a complete solution, because $e^{s_a t}$ becomes a factor of each term and, in effect, there is then only one arbitrary constant, which is not sufficient to accommodate two initial conditions. Treatment of that special case will be deferred until we develop some interpretive ideas about Eqs. (6-19). Referring to Eqs. (6-17) we see that if $\alpha \geqq 2\omega_0$ the eigenvalues are real, and that they are complex if $\alpha < 2\omega_0$. Discussion of the implications of these two conditions will be an important part of our results. Another useful interpretation involves the relationship of the analytic solution to the phase plane solution presented earlier.

Concerning the phase plane, Fig. 6-13 shows the initial state point of a typical problem, and the obvious graphical interpretation of Eqs. (6-20).

Although s_a and s_b can be complex, and they appear in Eq. (6-20), it is necessary that $x_2(0)$ must be real. This is possible if A_1 and B_1 are also complex. In fact, since s_b is the conjugate of s_a, it can be shown that B_1 must be the conjugate of A_1. This includes the possibility of A_1 and B_1 being equal and real, as a special case of the conjugate relationship, when the imaginary part is zero.[+] Thus, in general, A_1 and B_1 can be complex and they are related by

$$\bar{B}_1 = \bar{A}_1^*$$

Bars are used over the symbols to indicate the possibility of their being complex.

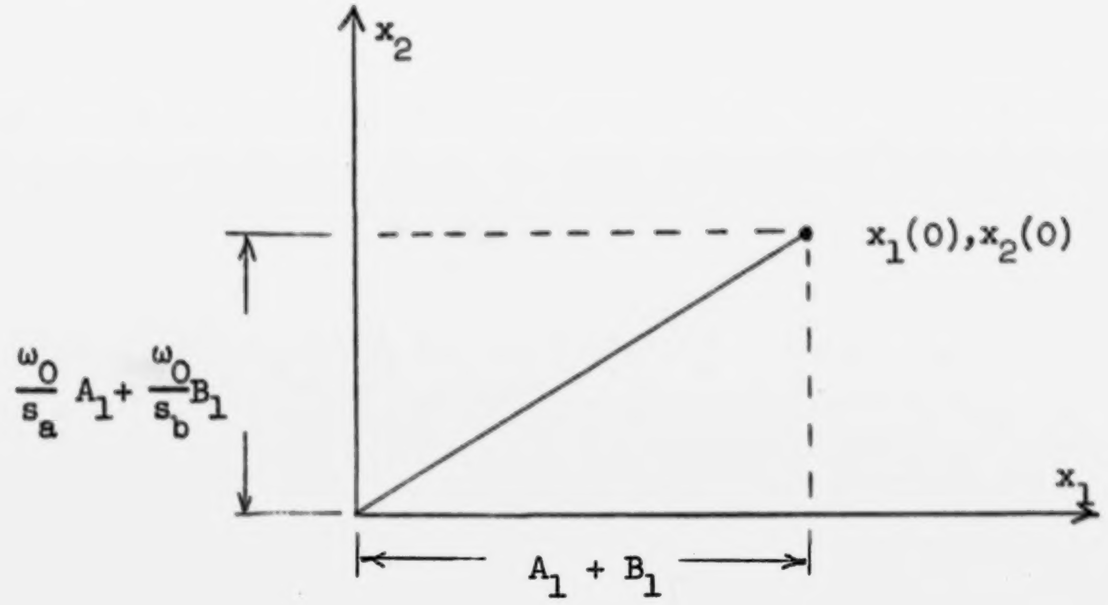


Figure 6-13

[+] To prove that $\bar{B}_1 = \bar{A}_1^*$ is necessary and sufficient for $x_2(0)$ to be real, adopt the notation $\bar{A}_1 = a_1 + ja_2$ and $\bar{B}_1 = b_1 + jb_2$. Also let $Me^{j\alpha} = (\omega_0/s_a)$, which implies that $(\omega_0/s_b) = Me^{-j\alpha}$. The equation for $x_1(0)$ is then

$$x_2(0) = M\left[(\cos\alpha + j\sin\alpha)(a_1 + ja_2) + (\cos\alpha - j\sin\alpha)(b_1 + jb_2)\right]$$

and its imaginary part is

$$M\left[a_1 \sin\alpha + a_2 \cos\alpha - b_1 \sin\alpha + b_2 \cos\alpha\right]$$

Setting this equal to zero gives

$$(a_2 + b_2)\cos\alpha = (a_1 - b_1)\sin\alpha$$

This equation must be true for any value of α. Therefore, the quantities in parentheses must be zero, and so

$$b_1 = a_1$$

$$b_2 = -a_2$$

Thus, $\bar{B}_1 = b_1 + jb_2 = a_1 - ja_2$, which is $\bar{A}^*$.

6-5 Interpretation of the Solutions

With the exception of the case $s_a = s_b$, Eqs. (6-20) can be solved in any numerical case. Thus, no additional analysis is needed to obtain numerical results for a specific problem. However, some general properties of the response curves can be determined. Also, the solution for the special case $s_a = s_b$ is yet to be determined. These are the topics to be considered in this section.

We shall treat three cases: (1) s_a and s_b are real; (2) $s_a = s_b$; (3) s_a and s_b are complex. Equations (6-19) and (6-20) serve as the starting point, the former providing the general form of the solution, and the latter containing the relationships from which A_1 and B_1 are determined.

Case (1): s_a and s_b are real.

Referring to Eqs. (6-17), it is seen that we can write

$$\left.\begin{aligned} s_a &= \sigma + \gamma \\ s_b &= \sigma - \gamma \end{aligned}\right. \qquad (6\text{-}21)$$

where

$$\left.\begin{aligned} \sigma &= -\frac{\alpha}{2} && \text{(a)} \\ \gamma &= \sqrt{\frac{\alpha^2}{4} - \omega_0^2} && \text{(b)} \end{aligned}\right\} \qquad (6\text{-}22)$$

By straightforward algebraic manipulation, it is found that

$$\left.\begin{aligned} A_1 &= \frac{1}{2\gamma}\Big[(\sigma+\gamma)x_1(0) - \omega_0 x_2(0)\Big] && \text{(a)} \\ B_1 &= \frac{1}{2\gamma}\Big[-(\sigma-\gamma)x_1(0) + \omega_0 x_2(0)\Big] && \text{(b)} \end{aligned}\right\} \qquad (6\text{-}23)$$

and

$$\left.\begin{aligned} A_2 &= \left(\frac{\omega_0}{\sigma+\gamma}\right)A_1 = \frac{1}{2\gamma}\Big[\omega_0 x_1(0) - (\sigma-\gamma)x_2(0)\Big] && \text{(a)} \\ B_2 &= \left(\frac{\omega_0}{\sigma-\gamma}\right)B_1 = \frac{1}{2\gamma}\Big[-\omega_0 x_1(0) + (\sigma+\gamma)x_2(0)\Big] && \text{(b)} \end{aligned}\right\} \qquad (6\text{-}24)$$

It is rather easy to check that these are solutions by observing that the sum of the first pair is $x_1(0)$, as required by Eq. (6-20), and the sum of the second pair is $x_2(0)$, in agreement with Eq. (6-20).

Using Eqs. (6-23) and (6-24), we are now interested in interpretations of

$$\begin{aligned} x_1 &= A_1 e^{(\sigma+\gamma)t} + B_1 e^{(\sigma-\gamma)t} \\ x_2 &= A_2 e^{(\sigma+\gamma)t} + B_2 e^{(\sigma-\gamma)t} \end{aligned} \qquad (6\text{-}25)$$

First consider the exponentials

$$e^{(\sigma+\gamma)t} \qquad \text{and} \qquad e^{(\sigma-\gamma)t}$$

observing that σ is negative, but greater in absolute value than γ. Thus, for all positive t,

$$(\sigma-\gamma)t < (\sigma+\gamma) < 0$$

and also

$$e^{(\sigma-\gamma)t} < e^{(\sigma+\gamma)t}$$

For a typical case, these functions are related as in Fig. 6-14, illustrating that $e^{(\sigma-\gamma)t}$ decays faster than $e^{(\sigma+\gamma)t}$. All solutions consist of suitable combinations of these, according to Eqs. (6-23), (6-24), (6-25).

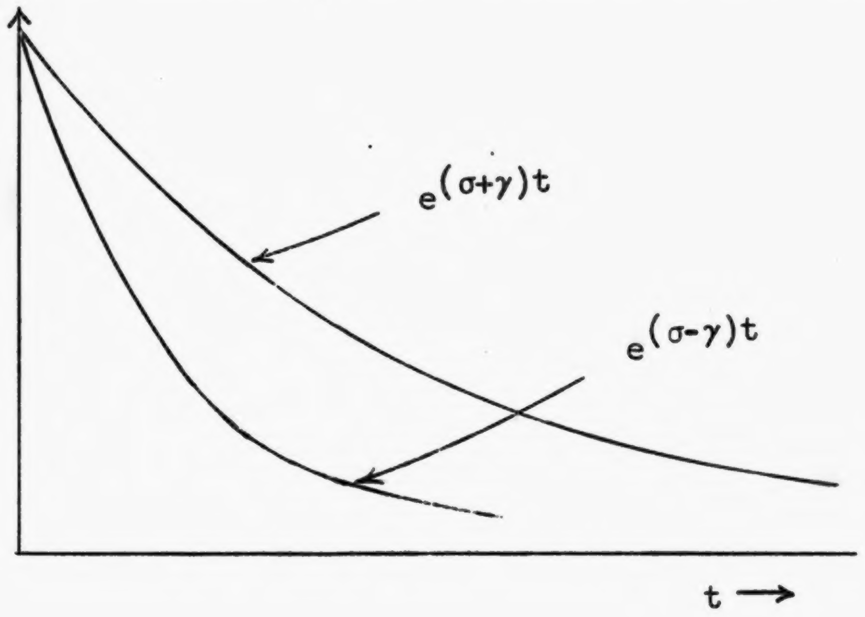


Figure 6-14.

[+]It would be convenient to have some general principles from which it would be possible to estimate response curve shapes from the initial conditions. It will be recalled that this was done in the case of R-C and R-L circuits, where the response consists of a single exponential. Those cases are handled by fitting a suitable exponential curve between initial and final values, where the parameters of the circuit determine the time constant of the exponential.

We would like to do the same thing in this case, but now there are two exponential functions, with time constants $-1/s_a$ and $-1/s_b$. This makes the problem more difficult, as illustrated by the three examples of Fig. 6-15. Depending on relative signs and sizes of A_1 and B_1, there are three possible curves. Thus, it is apparent that consideration of Eqs. (6-25) does not yield all the information we need. The reason is that both initial conditions affect the curve shape, whereas in looking at Fig. 6-15 only $x_1(0)$ is in evidence. The need to incorporate both initial conditions naturally suggests looking at the phase plane for the possible answer.

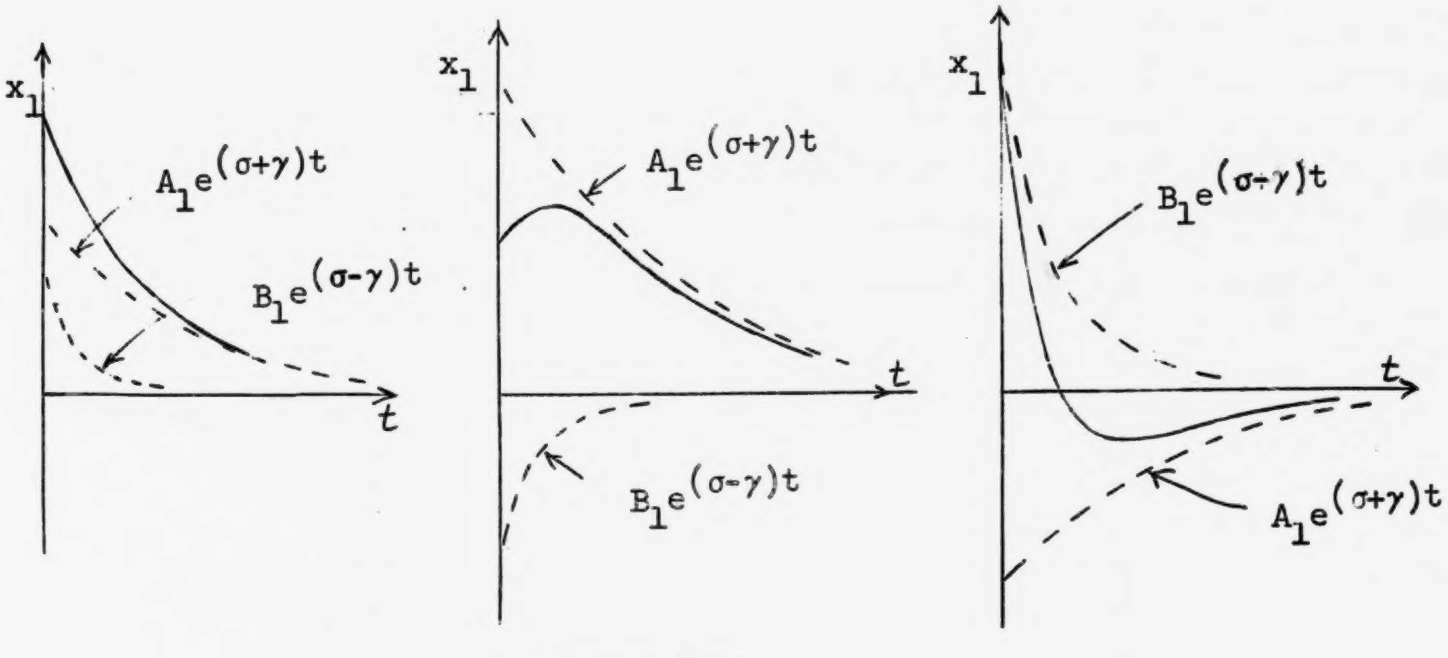


Figure 6-15.

There is an initial state point having coordinates $x_1(0)$ and $x_2(0)$, and a final state point at the origin. First, let us consider the behavior of the trajectory as t approaches infinity. To do this, consider the ratio x_2/x_1 as

---

[+]The remaining discussion of case (1) can be omitted without loss of continuity.

t approaches infinity. It is recalled that $e^{(\sigma-\gamma)t}$ approaches zero faster than $e^{(\sigma+\gamma)t}$. Therefore, for very large t we have (approximately)

$$\frac{x_2}{x_1} = \frac{A_2 e^{(\sigma+\gamma)t}}{A_1 e^{(\sigma+\gamma)t}} = \frac{A_2}{A_1}$$

This ratio is negative, and given by

$$\frac{A_2}{A_1} = \frac{\sigma-\gamma}{\omega_0} \tag{6-26$^+$}$$

In other words, the trajectory must approach the origin in such a way that the ratio x_2/x_1 will be nearly constant, which means the state point approaches the origin so as to be tangent to a radial line of slope $(\sigma-\gamma)/\omega_0$.

This result is portrayed in Fig. 6-16a. Some ambiguity remains, however. Equation (6-26) gives no information for making a choice among the four cases shown in Fig. 6-16a, all of which are tangent to the diagonal line at the origin.

Further information depends upon investigating properties of the trajectory at finite values of t. Help is obtained by returning to the original differential equations, to determine those points in the phase plane at which the velocity of the state point is toward the origin.

Referring to Fig. 6-17, it is seen that this will occur when

$$\frac{dx_2}{dx_1} = \frac{x_2}{x_1}$$

[+] From Eqs. (6-23) and (6-24)

$$\frac{A_2}{A_1} = \frac{\omega_0 x_1(0) - (\sigma-\gamma)x_2(0)}{(\sigma+\gamma)x_1(0) - \omega_0 x_2(0)} = \left(\frac{\sigma-\gamma}{\omega_0}\right)\left[\frac{\left(\frac{\omega_0}{\sigma-\gamma}\right)x_1(0) - x_2(0)}{\left(\frac{\sigma+\gamma}{\omega_0}\right)x_1(0) - x_2(0)}\right]$$

The quantity in brackets is unity, as may be seen by referring to Eq. (6-17) which shows that $\omega_0^2 = \sigma^2 - \gamma^2$, or $\omega_0/(\sigma-\gamma) = (\sigma+\gamma)/\omega_0$.

1 From Eqs. (6-5) and (6-6),

$$\frac{dx_2}{dx_1} = - \frac{\omega_0 x_1}{\alpha x_1 + \omega_0 x_2} = - \frac{\omega_0}{\alpha + \omega_0 \frac{x_2}{x_1}}$$

2 Thus, for the condition specified

$$- \frac{\omega_0}{\alpha + \omega_0 \frac{x_2}{x_1}} = \frac{x_2}{x_1}$$

or

3

$$\omega_0 \left(\frac{x_2}{x_1}\right)^2 + \alpha\left(\frac{x_2}{x_1}\right) + \omega_0 = 0$$

which has the two solutions

4

$$\frac{x_2}{x_1} = - \frac{\alpha}{2\omega_0} \pm \frac{1}{\omega_0} \sqrt{\frac{\alpha^2}{4} - \omega_0^2}$$

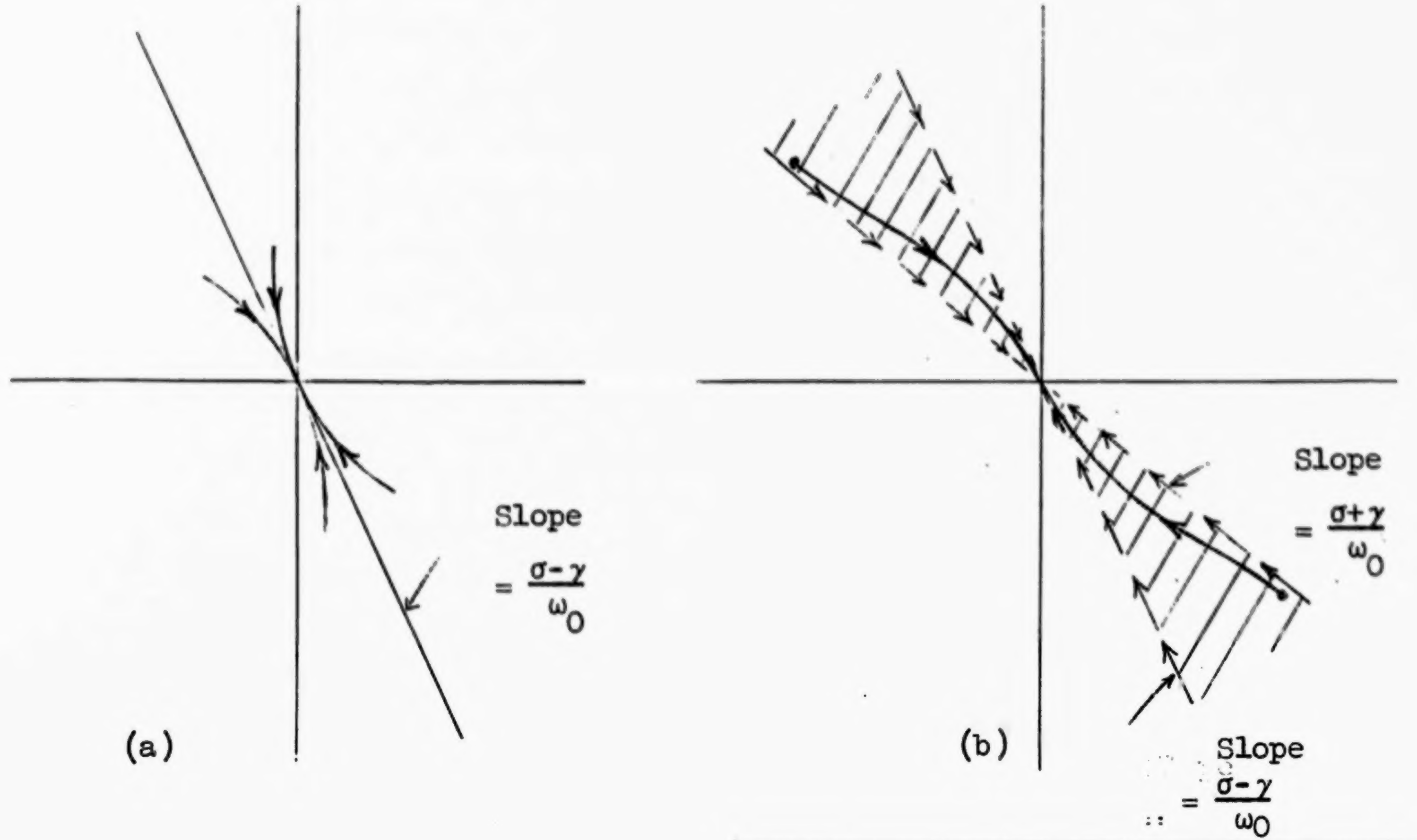
$$= \frac{\sigma+\gamma}{\omega_0} \quad \text{and} \quad \frac{\sigma-\gamma}{\omega_0}$$

5 These are constant values of the ratio x_2/x_1, which correspond to values on the radial straight lines shown in Fig. 6-16b. One of these is the line previously shown in Fig. 6-16a.

A trajectory cannot cross either of these lines. To do so would imply a contradiction, since it would require a velocity component normal to the line. 6 This observation leads to the further conclusion that for any initial state point lying within the shaded sectors, the entire trajectory must lie within the sector, and have a form of the general nature illustrated by the two examples in Fig. 6-16b.[+]

7 If the initial state point is in one of the regions shown shaded in Fig. 6-16c, it is clear from observation of the figure that the trajectories must be as indicated. All other possibilities would either cross one of the diagonal lines, or would not approach the origin tangentially to the heavy line.

[+] It can be shown that $dx_2/dx_1 > x_2/x_1$ at any point within the shaded sectors of Fig. 6-16b.


Slope
$= \frac{\sigma-\gamma}{\omega_0}$
(a)
Slope
$= \frac{\sigma+\gamma}{\omega_0}$
(b)
Slope
$= \frac{\sigma-\gamma}{\omega_0}$

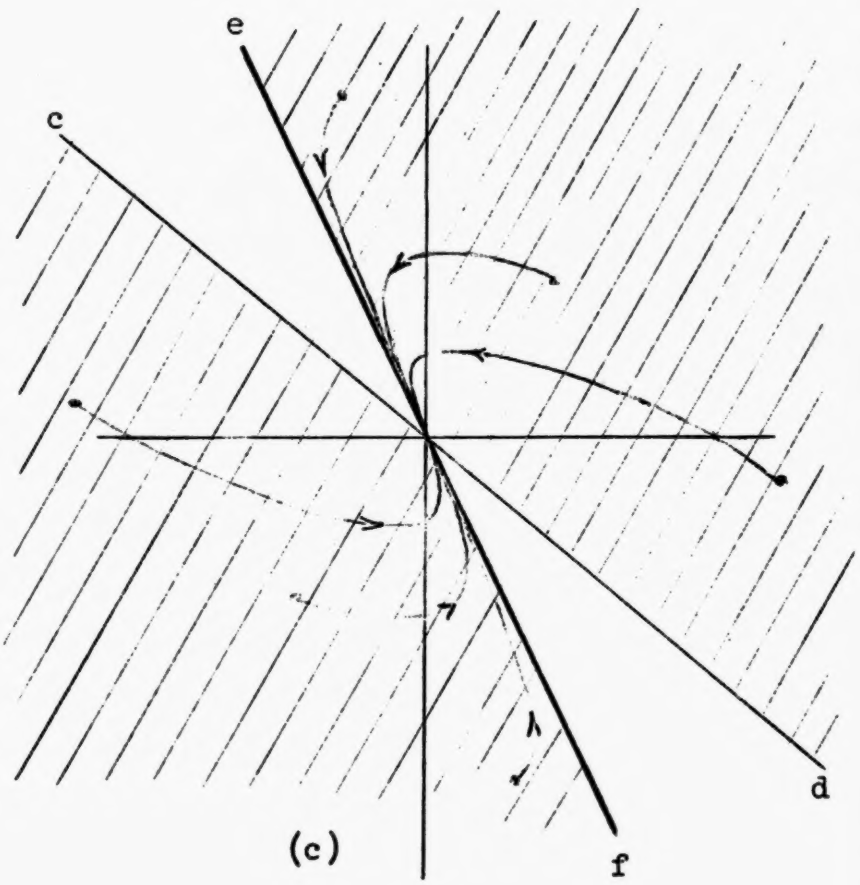


Figure 6-16.

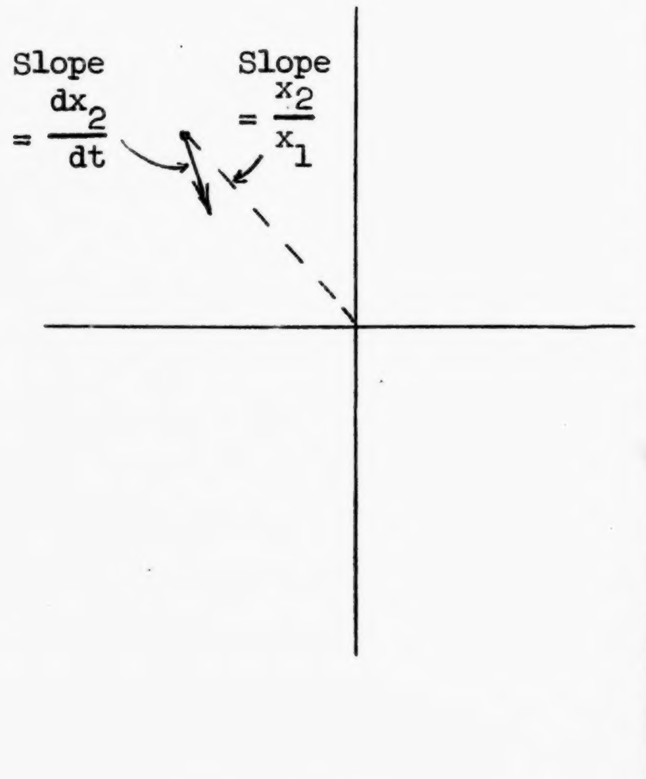

Slope
$= \frac{dx_2}{dt}$
Slope
$= \frac{x_2}{x_1}$

Figure 6-17.

Of course, this information does not determine the trajectory with precision. However, one further clue is obtained by observing that the trajectory must be normal to the x_2 axis as it crosses that axis. This is because $dx_2/dt = \omega_0 x_1 = 0$ at points on this axis.

These various trajectories can be used, in the customary manner of projecting onto the axes, to estimate shapes of the x_1 and x_2 curves. The existence of two sets of "forbidden" lines, which cannot be crossed by the trajectory gives a vivid confirmation that the responses obtained in the case of real eigenvalues cannot change sign more than once.

Case (2): $s_a = s_b$

The solution for this case is obtained from the previous one, by investigating the limit as γ approaches zero. To begin, the case (1) solutions are written explicitly, by multiplying A_1 and A_2 by $e^{(\sigma+\gamma)t}$, and B_1 and B_2 by $e^{(\sigma-\gamma)t}$, where the A's and B's are given by Eqs. (6-23) and (6-24). In the interest of brevity, let the factor $e^{\sigma t}$ be extracted while the remaining terms are combined with $x_1(0)$ and $x_2(0)$ as factors. The results are

$$x_1 = \frac{e^{\sigma t}}{2\gamma}\left\{ x_1(0)\left[(\sigma+\gamma)e^{\gamma t} - (\sigma-\gamma)e^{-\gamma t}\right] - x_2(0)\left[\omega_0 e^{\gamma t} - \omega_0 e^{-\gamma t}\right]\right\}$$

$$x_2 = \frac{e^{\sigma t}}{2\gamma}\left\{ x_1(0)\left[\omega_0 e^{\gamma t} - \omega_0{}^{-\gamma t}\right] - x_2(0)\left[(\sigma-\gamma)e^{\gamma t} - (\sigma+\gamma)e^{-\gamma t}\right]\right\}$$

or

$$x_1 = e^{\sigma t}\left[\sigma x_1(0)\left(\frac{e^{\gamma t}-e^{-\gamma t}}{2\gamma}\right) + x_1(0)\left(\frac{e^{\gamma t}+e^{-\gamma t}}{2}\right) - \omega_0 x_2(0)\left(\frac{e^{\gamma t}-e^{-\gamma t}}{2\gamma}\right)\right] \quad \text{(a)}$$

$$(6\text{-}27)$$

$$x_2 = e^{\sigma t}\left[\omega_0 x_1(0)\left(\frac{e^{\gamma t}-e^{-\gamma t}}{2\gamma}\right) - \sigma x_2(0)\left(\frac{e^{\gamma t}-e^{-\gamma t}}{2\gamma}\right) + x_2(0)\left(\frac{e^{\gamma t}+e^{-\gamma t}}{2}\right)\right] \quad \text{(b)}$$

Now let γ approach zero. The quantity $(e^{\gamma t}+e^{-\gamma t})/2$ approaches unity. However, $(e^{\gamma t}-e^{-\gamma t})/2\gamma$ becomes indeterminate. One way to obtain the limit is to use ℓ'Hospital's rule, differentiating numerator and denominator with respect to γ. The result is

$$\frac{te^{\gamma t}+te^{-\gamma t}}{2} = t$$

in the limit as γ goes to zero. Thus, when $\gamma = 0$, Eqs. (6-27) become

$$x_1 = e^{\sigma t}\left[\sigma t\, x_1(0) + x_1(0) - \omega_0 t\, x_2(0)\right]$$

$$x_2 = e^{\sigma t}\left[\omega_0 t\, x_1(0) - \sigma t x_2(0) + x_2(0)\right]$$

This can be further simplified, since in this case

$$\gamma = \sqrt{\frac{\alpha^2}{4} - \omega_0^2} = 0$$

Thus,

$$\omega_0 = \frac{\alpha}{2} = -\sigma$$

(rejecting $-\alpha/2$ because $\omega_0 = 1/\sqrt{LC}$ must be positive). The final solutions are

$$\begin{aligned} x_1 &= \left[x_1(0) + x_2(0)\right](\sigma t)e^{\sigma t} + x_1(0)e^{\sigma t} \qquad &(a)\\ x_2 &= \left[-x_1(0) + x_2(0)\right](\sigma t)e^{\sigma t} + x_2(0)e^{\sigma t} \qquad &(b)\end{aligned} \qquad (6\text{-}28)$$

Although the equations are different from those for case (1), the curves of x_1 and x_2 are similar. (The phase plane interpretation of Fig. 6-16 applies, but lines c-d and e-f now coincide, abolishing one of the regions.)+

Case (3): s_a and s_b are complex.

For this case, γ is imaginary. Therefore, it is convenient to change notation to

$$\begin{aligned} s_a &= \sigma + j\omega \\ s_b &= \sigma - j\omega \end{aligned} \qquad (6\text{-}29)$$

where

$$\omega = -j\gamma = \sqrt{\omega_0^2 - \frac{\alpha^2}{4}} \qquad (6\text{-}30)$$

+This statement refers to material which you may have omitted, in accordance with a previous footnote.

The algebra leading up to Eqs. (6-27) remains valid if γ is imaginary. Therefore, the required solutions are obtained immediately by replacing γ by $j\omega$, giving

$$x = e^{\sigma t}\left[\sigma x_1(0)\left(\frac{e^{j\omega t}-e^{-j\omega t}}{2j\omega}\right) + x_1(0)\left(\frac{e^{j\omega t}+e^{-j\omega t}}{2}\right) - \omega_0 x_2(0)\left(\frac{e^{j\omega t}-e^{-j\omega t}}{2j\omega}\right)\right]$$

$$x_2 = e^{\sigma t}\left[\omega_0 x_1(0)\left(\frac{e^{j\omega t}-e^{-j\omega t}}{2j\omega}\right) - \sigma x_2(0)\left(\frac{e^{j\omega t}-e^{-j\omega t}}{2j\omega}\right) + x_2(0)\left(\frac{e^{j\omega t}+e^{-j\omega t}}{2}\right)\right]$$

However, the exponential combinations represent trigonometric functions, as follows:

$$\frac{e^{j\omega t}-e^{-j\omega t}}{2j} = \sin \omega t$$

$$\frac{e^{j\omega t}+e^{-j\omega t}}{2} = \cos \omega t$$

Thus, the solutions become

$$x_1 = \left\{\left[\frac{\sigma x_1(0)-\omega_0 x_2(0)}{\omega}\right]\sin \omega t + x_1(0)\cos \omega t\right\} e^{\sigma t} \qquad \text{(a)}$$

$$\text{(6-31)}$$

$$x_2 = \left\{\left[\frac{\omega_0 x_1(0)-\sigma x_2(0)}{\omega}\right]\sin \omega t + x_2(0)\cos \omega t\right\} e^{\sigma t} \qquad \text{(b)}$$

The two trigonometric functions within the brackets combine to yield sinusoidal waves of angular frequency ω. It is recalled from Eq. (6-31) that

$$\omega = \sqrt{\omega_0^2 - \frac{\alpha^2}{4}}$$

but, since $\sigma = -\alpha/2$, this is also

$$\omega = \sqrt{\omega_0^2 - \sigma^2} = \omega_0\sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2} \qquad \text{(6-32)}$$

This is an interesting result, giving an analytic proof of the fact mentioned in connection with Fig. 6-7, that ω, the natural angular frequency of the wave, is less than ω_0.

For interpretation of these equations, it is convenient to combine the trigonometric functions. The quantities in brackets are steady state sinusoids, to which phasor ideas apply, with the results+

$$x_1 = X_0 \cos\left[\omega_0\sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\; t - \theta_1\right] e^{\sigma t} \qquad \text{(a)}$$

$$x_2 = X_0 \cos\left[\omega_0\sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\; t - \theta_2\right] e^{\sigma t} \qquad \text{(b)} \qquad (6\text{-}33)$$

where

$$X_0 = \sqrt{\frac{\omega_0^2\left\{\left[x_1(0)\right]^2 + \left[x_2(0)\right]^2\right\} - 2\omega_0\sigma x_1(0)x_2(0)}{\omega_0^2 - \sigma^2}} \qquad (6\text{-}34)$$

$$\theta_1 = \arctan\left[\frac{\sigma - \omega_0\dfrac{x_2(0)}{x_1(0)}}{\sqrt{\omega_0^2 - \sigma^2}}\right] \qquad \text{(a)}$$

$$\theta_2 = \arctan\left[\frac{\omega_0\dfrac{x_1(0)}{x_2(0)} - \sigma}{\sqrt{\omega_0^2 - \sigma^2}}\right] \qquad \text{(b)} \qquad (6\text{-}35)$$

+Using x_1 as an example, since $\sin \omega t = \cos(\omega t - \frac{\pi}{2})$, the phasors for the $\sin \omega t$ and $\cos \omega t$ terms are, respectively,

$$-j\left[\frac{\sigma x_1(0) - \omega_0 x_2(0)}{\omega}\right] \quad \text{and} \quad x_1(0)$$

Accordingly, the phasor for the sum is

$$\sqrt{x_1(0)^2 + \frac{\sigma^2\left[x_1(0)\right]^2 - 2\sigma\omega_0 x_1(0)x_2(0) + \omega_0^2\left[x_2(0)\right]^2}{\omega 2}}\;\angle\; -\arctan\left[\frac{\sigma - \omega_0\dfrac{x_2(0)}{x_1(0)}}{\omega}\right]$$

$$= \sqrt{\frac{\omega_0^2\left\{\left[x_1(0)\right]^2 + \left[x_2(0)\right]^2\right\} - 2\sigma\omega_0 x_1(0)x_2(0)}{\omega_0^2 - \sigma^2}}\;\angle\; -\arctan\frac{\sigma - \omega_0\dfrac{x_2(0)}{x_1(0)}}{\sqrt{\omega_0^2 - \sigma^2}}$$

Both waves are of the general form shown in Fig. (6-18). Their amplitudes are the same, but their initial angles are different. The effect of the exponential in reducing the magnitudes of successive peaks is evident in the figure.

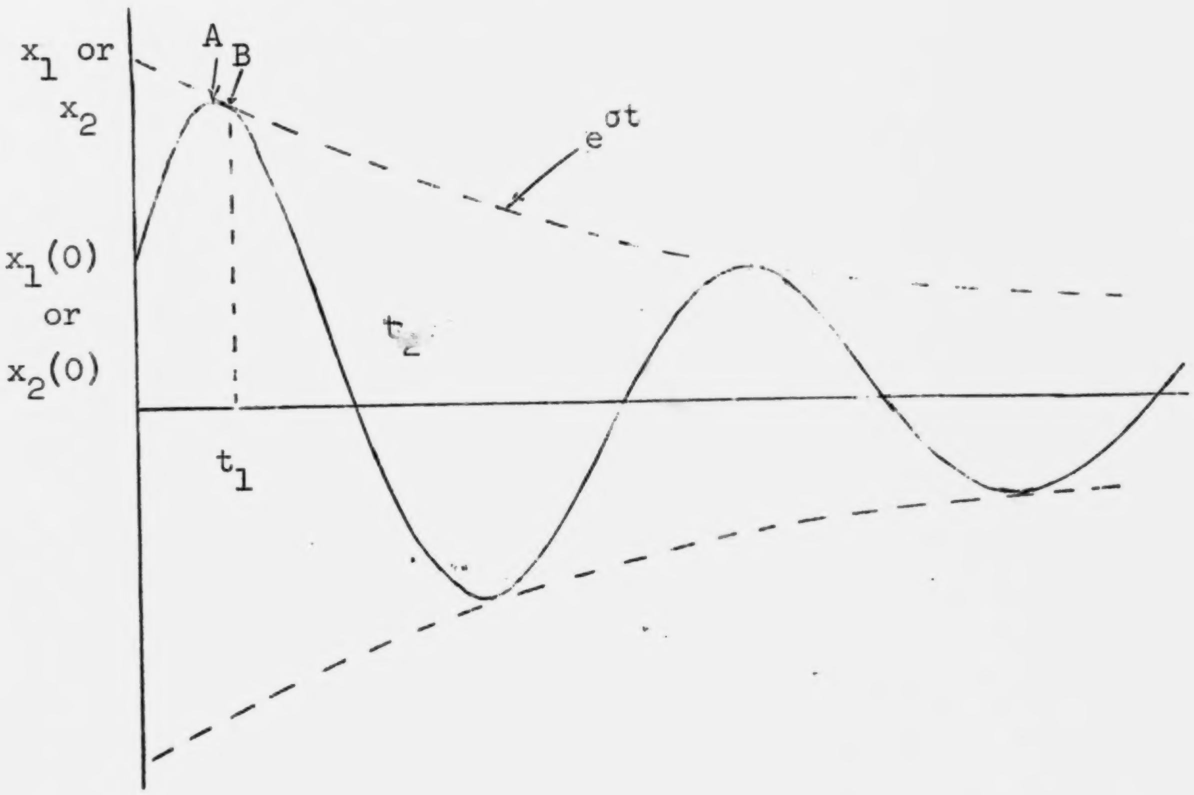


Figure 6-18.

By differentiating the equations it can be shown that points where the undamped sinusoid is maximum correspond to points (like B in Fig. 6-18) where the damped wave is tangent to the exponential envelope. The actual peak values (like A) occur slightly earlier. The separation between these two points becomes larger as the rate of decay increases. The first tangent point (B) occurs when the argument of the cosine is zero; that is, when $t = t_1$ is a solution of

$$\omega_0 \sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\, t_1 = \theta_1 \qquad \text{or} \qquad \omega_0 \sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\, t_1 = \theta_2$$

respectively, for the x_1 and x_2 curves. Also, the first zero occurs at t_2, where

$$\omega_0 \sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\, t_2 = \theta_1 + \frac{\pi}{2} \qquad \text{or} \qquad \omega_0 \sqrt{1-\left(\frac{\sigma}{\omega_0}\right)^2}\, t_2 = \theta_2 + \frac{\pi}{2}$$

1 With the aid of the envelope curves, and the reference points indicated in Fig. 6-18, it is relatively easy to sketch the response curves obtained for given initial conditions. However, it is well to recall the simplicity of viewpoint provided by the phase plane. Sketches like those in Fig. 6-19 quickly provide information about the x_1 and x_2 curves for various positions of the initial state point. 2

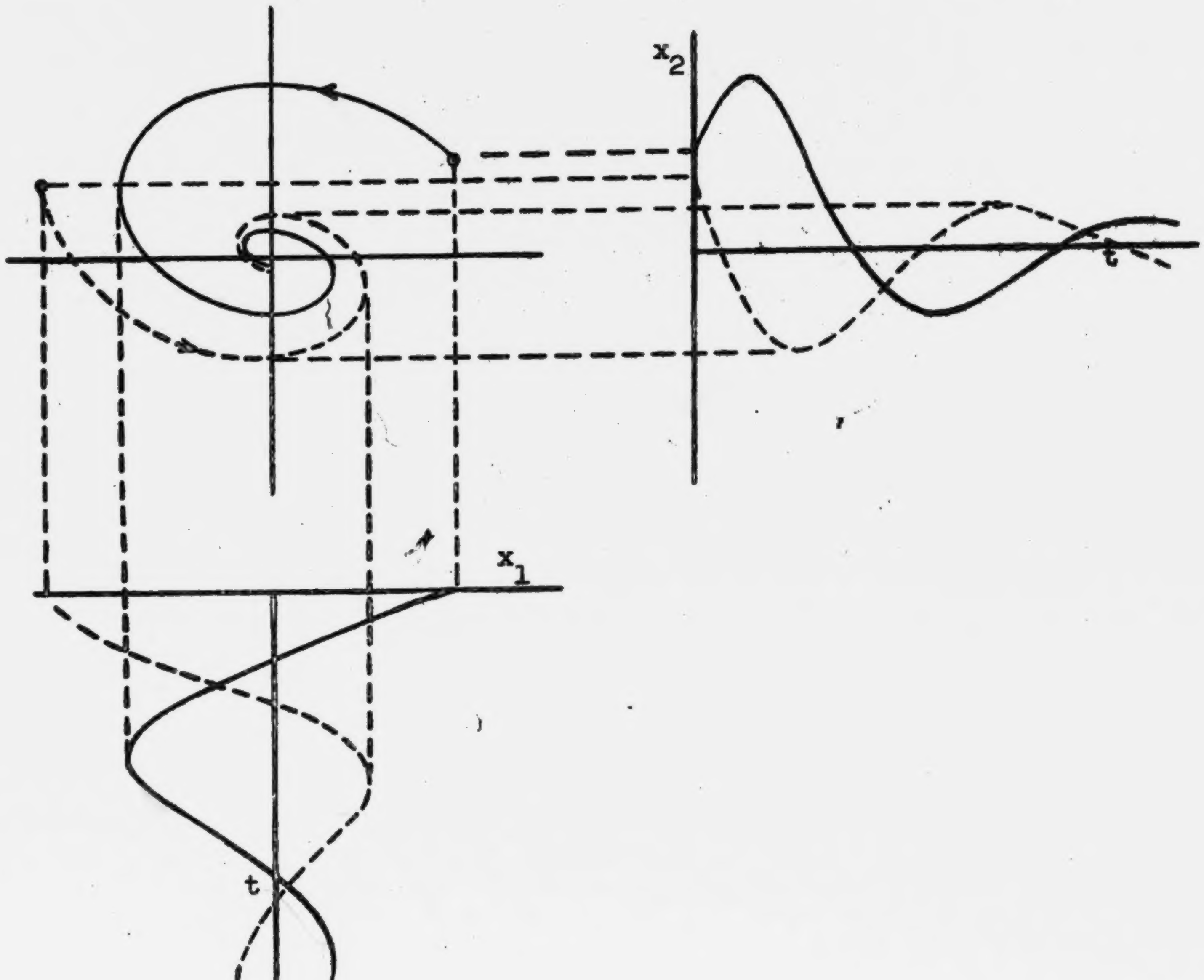


Fig. 6-19

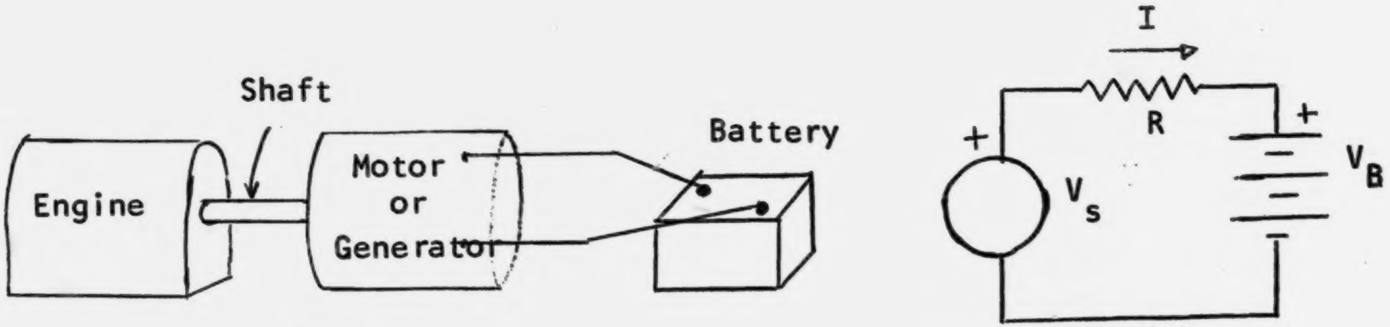
Chapter 8

ELECTRIC MOTORS

Introduction:

It is generally true that any electrical machine that can operate as a generator can also operate as a motor.* When such a machine is connected in a circuit and mechanically connected to a mechanical load, or possible source of driving torque, whether the machine operates as a motor or generator depends on conditions external to it. For example, the electric machine (motor or generator) attached to the axle of an electric locomotive is acting as a motor when it is receiving electric power from overhead lines and delivering mechanical power to drive the locomotive up a hill. Conversely, if the locomotive is going down hill, the direction of power flow through the machine can be reversed. It can be driven mechanically, and deliver electrical power to the overhead line. No electrical connections are changed in the process. Whether gravity is pulling back or forward determines the direction of the mechanical torque at the machine shaft, and hence the direction of power flow.

These principles are most easily seen in the case of a d-c machine. In Fig. 8-1a, a d-c machine with separately excited field (to make the flux independent of speed) is connected to a driving engine of some sort, which can drive the machine at various speeds, or be driven by it.



(a) (b)

Figure 8-1.

The equivalent circuit of the electric machine is shown in Fig. 8-1b. Resistor R includes the resistance in the equivalent circuits of both the generator and the battery, and also the connecting wires. The potential source V_s is equal the induced emf of the machine, and is proportional to its rotational speed ω_m, thus:

$$V_s = k\,\omega_m \tag{8-1}$$

where k is a constant of proportionality.

---

*An exception to this statement is discussed in Sec. 8-6.

Solution of Fig. 8-1b readily yeilds

$$I = \frac{k\,\omega_m - V_B}{R} \qquad (8\text{-}2)$$

If I is positive, electrical power leaves the machine and enters the battery. That is to say, the machine acts as a generator. If I is negative, electrical power leaves the battery and enters the machine, and so it is then acting as a motor.

If we use the example of the locomotive, suppose it is going up a hill at a speed ω_{m1}. The electric machine must be acting as a motor, and so in order to have a negative I, we must have

$$V_B > k\,\omega_{m1}$$

This condition can be accomplished by adjusting the supply voltage V_B, or by adjusting the field strength (which affects k) to make $k\omega_{m1}$ sufficiently small. Now, after passing the top of a hill, in beginning to coast down, ω_{m2} will increase. At any speed ω_{m2} where

$$V_B < k\,\omega_{m2}$$

the machine will act as a generator, because the current will have reversed.

8-1. Electromotive Force of a d-c Machine

As indicated in the introduction, the conditions relevant to induced emf depend only on the motion of armature conductors in a magnetic field. Such motion is present in either a motor or generator, and so the principles which govern induced emf are the same for both.

As you have seen in the consideration of induced emf (see programmed text) the emf depends on the rotational speed, the distribution of flux around the airgap, and the number of turns of wire on the armature, and the number of commutator segments. For a given machine, the armature winding is constant, and the flux distribution can be assumed to be reasonably constant, although the total flux distribution may vary slightly.

[illegible]

If E is the emf of a given machine, it is therefore true that it is proportional to speed and to the airgap flux Φ, and thus can be written

$$V_s = E = k_E \Phi \, \omega_m \tag{8-3}$$

where k_E is a proportionality constant which will vary from machine to machine. (What does k_E depend upon?)

Referring to Fig. 8-2, the flux Φ in this expression is the total flux through a plane P-P' located as indicated in the figure. For the flux and rotational directions shown in Fig. 8-2, the emf will be into the paper (indicated by the X symbol) for all conductors to the right of P-P', and will be out of the paper (indicated by the · symbol) for all conductors to the left of P-P'. The armature connections are such that half of these conductors are in series between the brushes, to produce emf E. Note that no reference direction is shown on the diagram for E, because the end connections of the armature conductors and commutator are not shown, making it impossible to define a reference direction for E for this particular diagram.

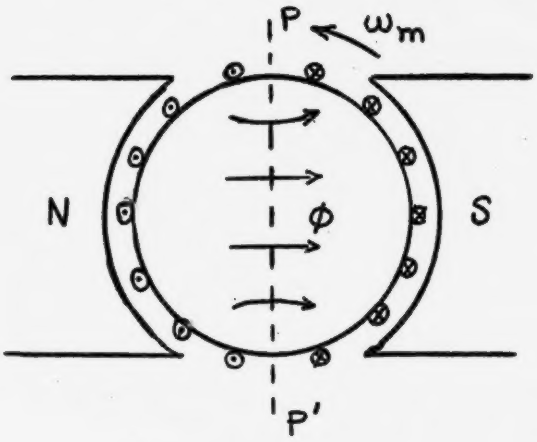


Figure 8-2.

8-2 Torque of a d-c Machine

When the machine portrayed symbolically in Fig. 8-2 acts as a generator, the currents in the armature conductors are in the same directions as the emfs, as indicated in that figure. If the machine is acting as a motor, the currents are reversed, as discussed in the introduction. Thus, Fig. 8-3a would represent relative directions of rotation, flux, and current, for a motor. The current reference directions shown will be for armature current I_a, which is also shown on the equivalent circuit in Fig. 8-3b. Thus, motor action will be associated with positive values of I_a. (This is opposite to the reference direction in Fig. 8-1, the change having been made because we want to concentrate on operation of a motor. By making this change we will avoid

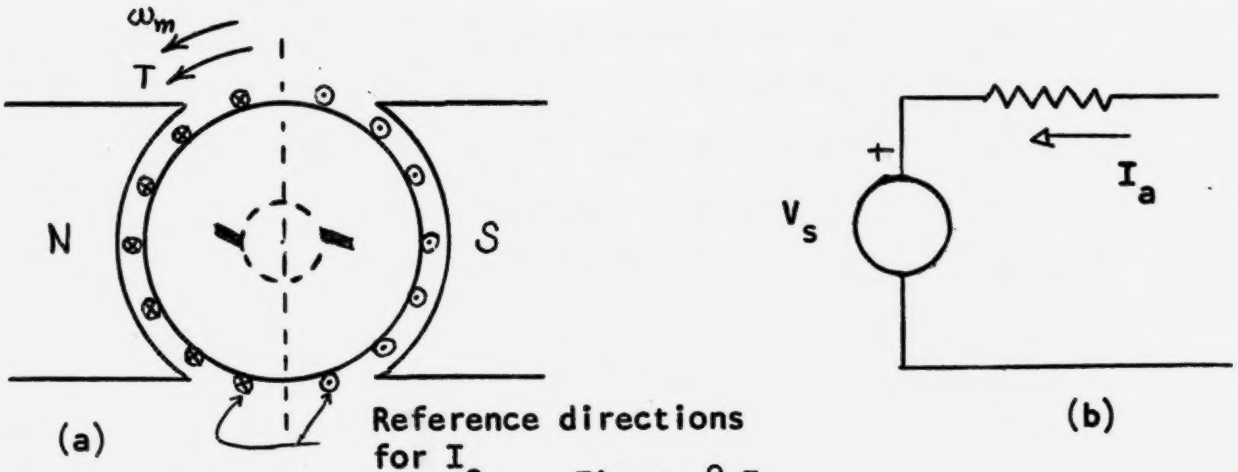

ω_m
T
N
S
(a)
Reference directions for I_a
V_s
+
I_a
(b)

Figure 8-3.

having always to consider I_a as negative. However, this change was quite arbitrary, and was unnecessary.)

The current in each armature conductor will be $I_a/2$ (Why?). Thus, considering a pair of diametrically opposite conductors, as in Fig. 8-4, each will experience a force as indicated. Specifically, if B_{m1} is the

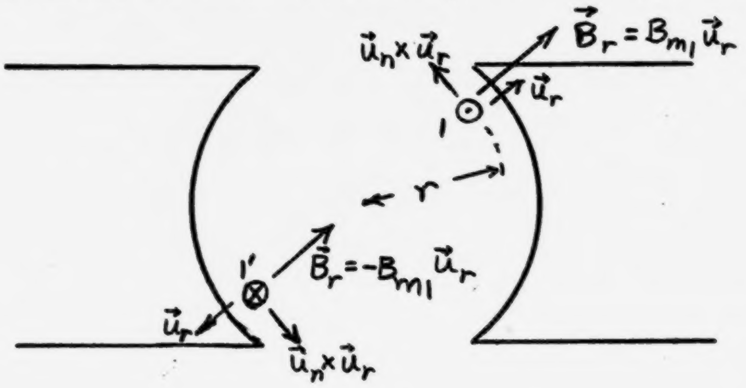

$\vec{u}_n \times \vec{u}_r$
$\vec{B}_r = B_{m1}\vec{u}_r$
$\vec{u}_r$
1
r
1'
$\vec{B}_r = -B_{m1}\vec{u}_r$
$\vec{u}_r$
$\vec{u}_n \times \vec{u}_r$

Figure 8-4.

magnitude of flux density at each point, $\vec{B}_r$ has the values indicated in the figure. Let $\vec{u}_n$ be a unit vector out of the paper. The force on conductor (1) will be

$$F_1 = \frac{B_{m1} I_a \ell}{2} (\vec{u}_n \times \vec{u}_r)$$

and the force on conductor (1') will be

$$F_{1'} = \frac{(-B_{m1})(-I_a)\ell}{2} (\vec{u}_n \times \vec{u}_r)$$

$$= \frac{B_{m1} I_a \ell}{2} (\vec{u}_n \times \vec{u}_r)$$

The directions of the unit vector products are indicated in the figure. It is evident that these forces are additive in producing torque about the axis of rotation, giving a torque

$$T_{11'} = B_{m'} I_a \ell r \qquad (8\text{-}4)$$

Each pair of conductors will produce a similar effect, and so the total torque will be a sum

$$T = T_{11'} + T_{12'} + \ldots \text{ etc.}$$

$$= I_a r \ell \, (Bm_1 + B_{m2} + \ldots)$$

The sum in parentheses is proportional to the total flux Φ passing through the rotor. (Why is it proportional rather than equal to the flux?) Thus, we can absorb r and ℓ in proportionality constant which depends on the machine's dimensions, to yield

$$T = k_T I_a \Phi \qquad (8\text{-}5)$$

as a relation that shows how torque varies with the essential variables I_a and Φ.

T is a scalar quantity in this expression, having the reference direction shown. That is, when T is positive it will be in the direction of rotation, confirming the earlier finding that positive values of I_a will correspond to motor action.

8-3 Speed-Torque Curves

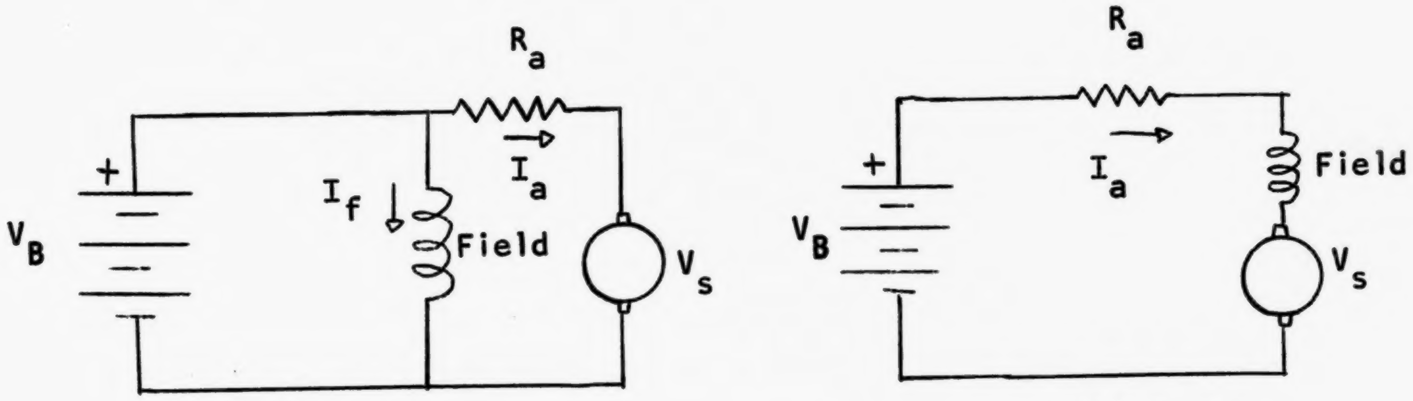
Much information about the operation of motors is obtained by a consideration of how speed varies with torque. This is because a mechanical load is characterized by a driving torque which also is a function of speed. Thus, if speed torque curves of a motor and a load are plotted on the same set of axes, they will intersect at the operating point.

The two equations

$$V_s = k_E \phi \omega_m \qquad \text{(a)}$$

$$T = k_T \phi I_a \qquad \text{(b)} \qquad (8\text{-}6)$$

and the equivalent circuits shown in Fig. 8-5 provide the information needed to predict how speed of a motor will vary with torque.



(a) Shunt

(b) Series

Figure 8-5.

8-4 Shunt Motor

Figure 8-5a illustrates a shunt motor, the word "shunt" referring to the fact that the field coil is "shunted across" the armature. The field current can be adjusted by means of a field rheostat R_f (a means of speed control, as we shall see) but in the absence of such an adjustment, ϕ remains constant.

From the equivalent circuit

$$V_s = V_B - I_a R_a$$

or

$$k_E \phi \omega_m = V_B - I_a R_a$$

1 Also, from Eq. (8-6b), $I_a = T/k_T\phi$ giving

$$k_E\phi\omega_m = V_B - (\frac{R_a}{k_T\phi})T$$

2 or

$$\omega_m = \frac{V_B}{k_E\phi} - (\frac{R_a}{k_E k_T\phi^2})T \qquad (8\text{-}7)$$

A set of ω_m vs. T curves for several values of ϕ (or I_f) is shown in Fig.
3 8-6.* Curves of I_a vs. T are also shown by the dashed lines. Labels 1, 2, and 3 indicate progressively increasing values of ϕ (or increasing I_f).

4

5

6

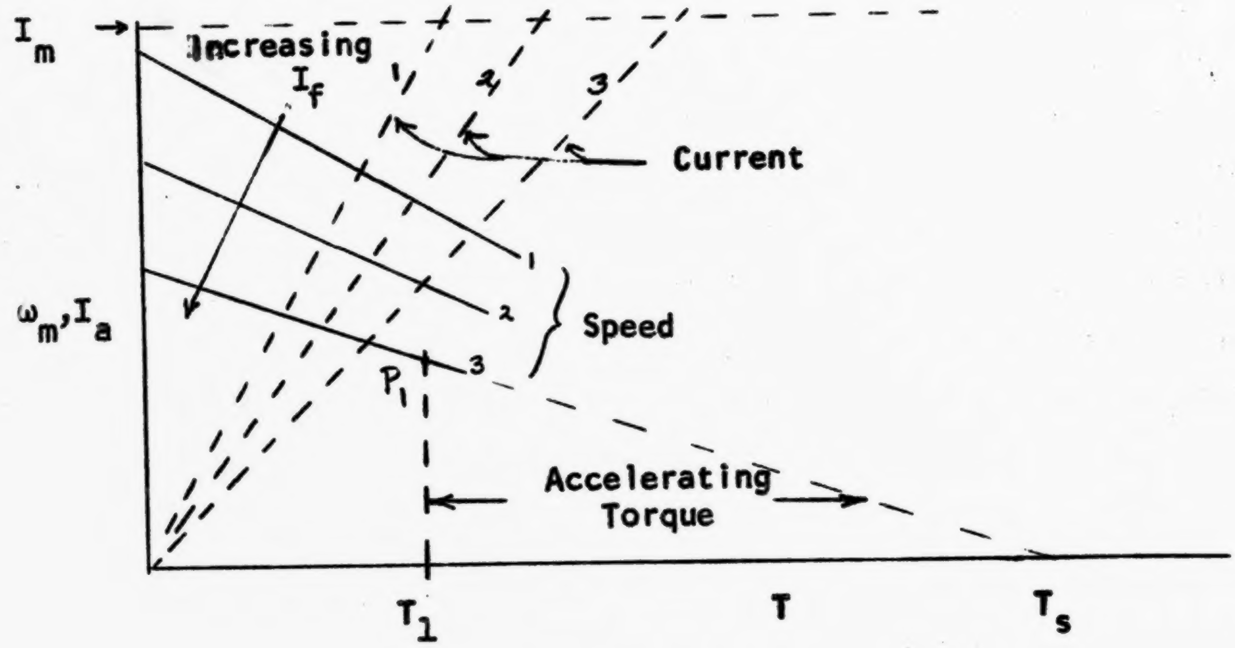


Figure 8-6.

7 Much can be learned from these curves about the operation of a shunt motor. Suppose it is required to start such a motor when its shaft is connected to a load for which the torque is independent of speed (say starting a locomotive on a hill), and that the field strength corresponds to curve 3.
8 The constant torque is represented by the dashed vertical line through T_1. Since the initial speed is zero, the initial torque provided by the motor will be the intercept T_s. The difference (T_s-T_1) is available for acceleration. This would seem to be satisfactory. With increasing speed there would gradually
9 be less accelerating torque, as indicated in the figure, and steady speed would be reached at point P_1 However, for any motor there is a value of I_a which

---

*The straight-line curves assume ϕ is independent of I_a, which is not exactly true due to armature reaction (see programmed text).

should not be exceeded, to avoid damage due to overheating. Suppose I_m is such a value. From the dashed curve (3) it is evident that the current limit would be exceeded at the start. Thus, although the motor would run satisfactorily at point P_1 under this specified load, it would not start properly.

Starting can be accomplished within the current limitation by temporarily inserting additional resistance in the armature circuit. As an explanation, from Eq. (7) it is seen that increasing R_a will increase the slope of the curve, as shown by the curve labeled "increased R_a" in Fig. 8-7. With this increased R_a, steady speed will be reached at P_2. If the added resistance is then removed, operation will shift to the original curve, and the motor torque will jump to T_0, and so the difference T_0-T_1 is available for acceleration. Acceleration will stop at point P_1.

This is a normal procedure for starting a shunt motor. As shown in Fig. 8-8, a high current-capacity resistor is provided with a series of taps which are contacted sequentially by a sliding arm. The number of taps depends on the characteristics of the load. For example, whereas one step was sufficient for a load torque T_1, for a greater torque, say T_2, two steps would be required.

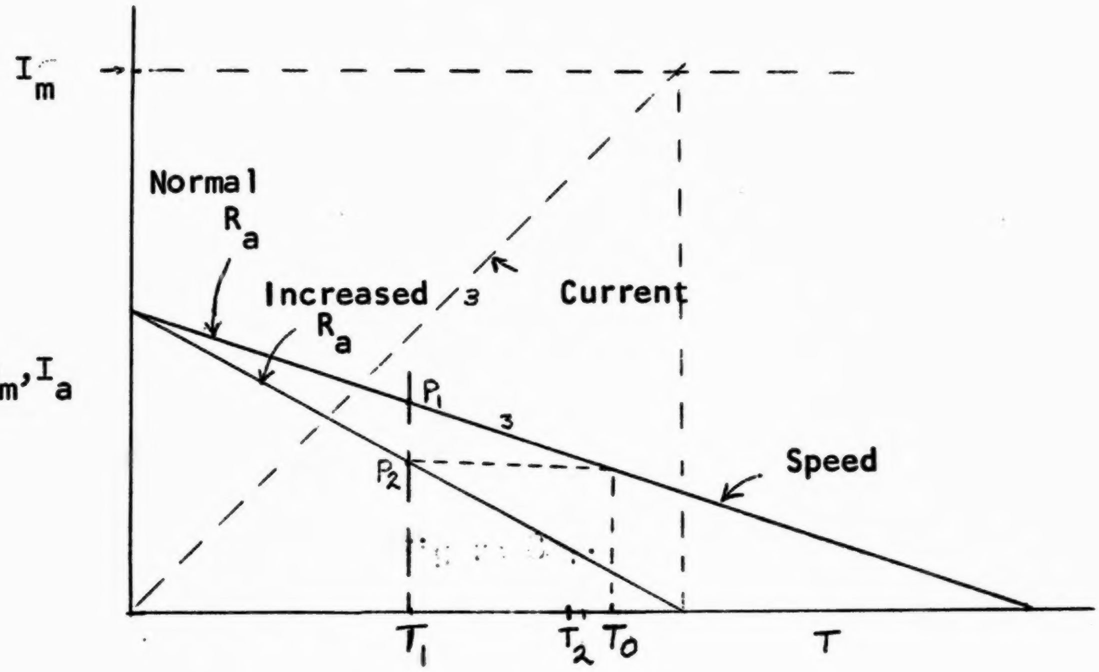


Figure 8-7.

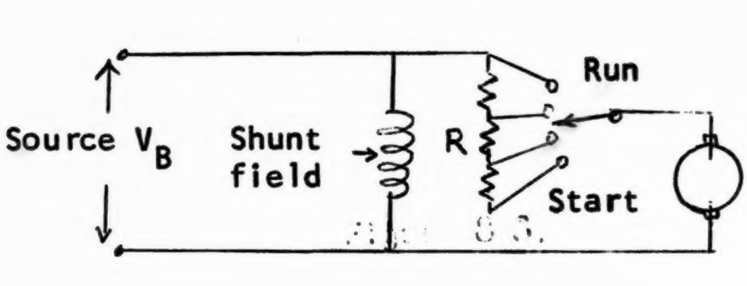


Figure 8-8

A still larger number of steps would be required to start along curves (2) or (1) of Fig. 8-6, which correspond to a weaker field. (Do you see why?)

In reference to Fig. 8-8, it should be observed that the field coil is directly across the source. That is, the starting resistor is not placed in the line to the motor. (Why not?)

By using the example of a constant torque (independent of speed) we have seen how the operating point at constant speed is determined by the intersection of the speed torque curves of the motor and the load. The constant torque curve (vertical line) typifies situations where the load torque is due to the pull of gravity, as in the previously cited example of a locomotive on a constant slope incline, or a building elevator.

Of course, friction always provides some of the load torque, and this is not independent of speed. Dry friction is nearly constant, but decreases somewhat after motion starts. A fan is a totally different type of load, having a torque which is nearly proportional to the square of speed. These three examples are illustrated in Fig. 8-9.

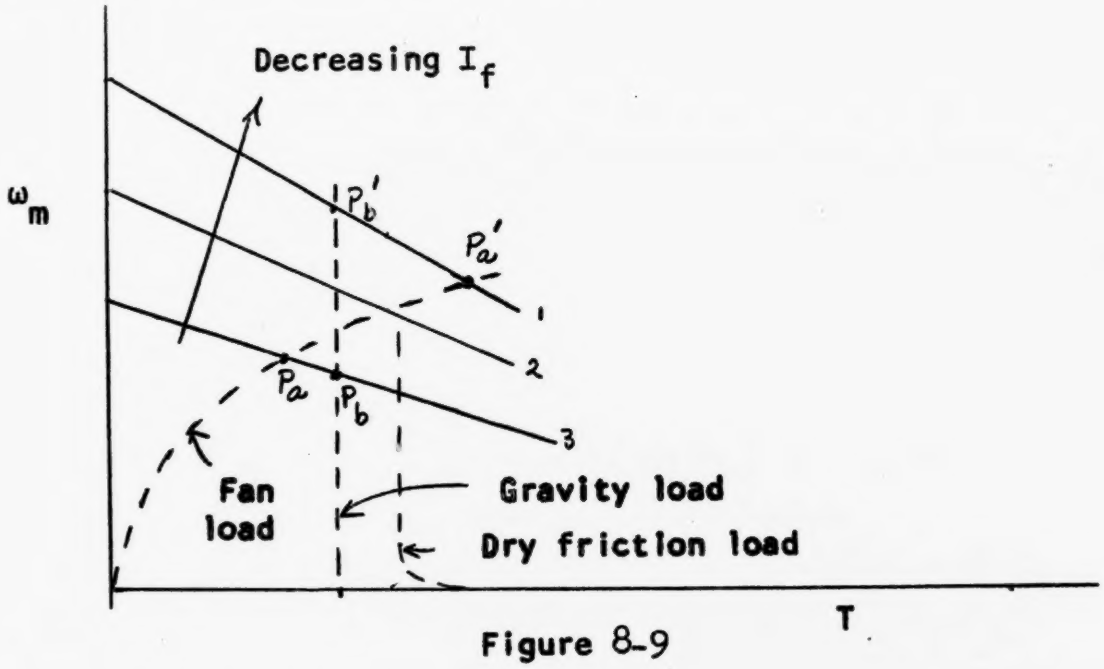


Figure 8-9

The effect of using a field current adjustment to increase speed is seen to be related to the characteristic of the load. For example, on curve (3) the fan and gravity loads will be driven at constant speeds represented respectively by points P_a and P_b. If the field current is decreased, shifting the motor characteristic to curve (1), operation will be respectively at points P_a' and P_b'.

The speed of the fan will increase less than the speed of the gravity load, because the torque requirement increases with speed.

The characteristic curves of shunt motors display the property of relatively small variation of speed over the operating range, for fixed field current. They are therefore useful in driving loads which should maintain nearly constant speed.

On the other hand, shunt motors have the disadvantage of not being efficient for starting loads which have high torque demands at low speed, such as gravity loads. As an explanation of this, refer to Fig. 8-10, which includes the motor and load speed-torque curves, the I_a vs. torque curve, with the indication of the maximum safe starting current I_m. Suppose this maximum starting current is twice the rated current for operation at constant load, which we shall call I_r. Furthermore, suppose the load torque is such that at constant speed the motor will operate at rated power output, which means that I_r will be the value of I_a. This constant speed operating condition is indicated by point P in Fig. 8-10.

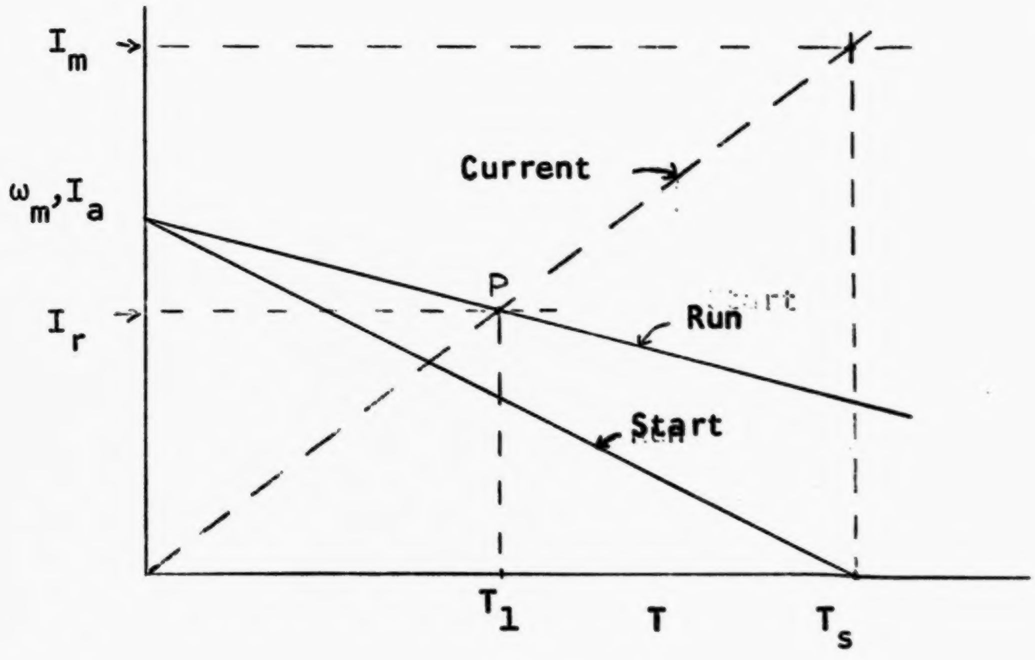


Figure 8-10.

We are interested in the maximum starting torque. This is T_s, which is twice T_1, as a result of torque being proportional to current. Thus, for this particular case, the torque available for acceleration is $T_s - T_1 = T_1$. This amount of accelerating torque might be unduly small, resulting in an undesirably slow start. For loads having small starting torques as, for example, a fan, the small accelerating torque would be less of a disadvantage. In summary, we have shown that for a shunt motor starting up under a constant torque load, near its maximum torque (as determined by I_r as the maximum armature current), the torque available for acceleration will be equal to this maximum torque.

8-5 Series Motor

If the field excitation is obtained by connecting a field coil in series with the armature, as in Fig. 8-11, the resulting machine is called a series motor. Compared to the shunt case, the series field coil will have a relatively small number of turns, because the armature current can be large compared with the current in a shunt coil.

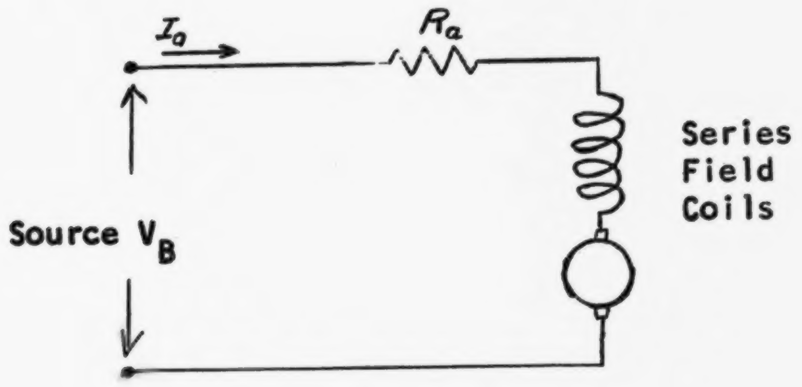


Figure 8-11.

The general properties of the speed-torque curves can be obtained from the two basic equations

$$k_E \phi \omega_m = V_B - I_a R_a$$

$$T = k_T \phi I_a$$

by making ϕ proportional to I_a. To do this is somewhat of an approximation, because in reality ϕ is related to I_a in accordance with a nonlinear magnetization curve. However, in the interest of simplicity, and since we are not interested in quantitative details, we shall use the linear approximation

$$\phi = k_\phi I_a \tag{8-8}$$

where k_ϕ is a constant which depends upon the number of turns on the field coil, and dimensions of the magnetic circuit.

Substituting Eq. (8-8) for Φ in the above equations yields

$$\begin{aligned} (k_E k_\Phi) I_a \omega_m &= V_B - I_a R_a & \text{(a)} \\ T &= (k_T k_\Phi) I_a^{\,2} & \text{(b)} \end{aligned} \qquad (8\text{-}9)$$

For a given machine, the quantities in parentheses are constants. Equation (8-9a) can be written

$$(k_E k_\Phi)\omega_m = \frac{V_B}{I_a} - R_a$$

and from Eq. (8-9b), $I_a = \sqrt{T/(k_T k_\Phi)}$. Therefore,

$$k_E k_\Phi \omega_m = \sqrt{\frac{k_T k_\Phi}{T}}\, V_B - R_a$$

or

$$\omega_m = \sqrt{\frac{k_T}{k_E^2 k_\Phi}}\, \frac{V_B}{\sqrt{T}} - \frac{R_a}{k_E k_\Phi}$$

Of course, since the field coils are in series with the armature, in the above equations R_a is the sum of the armature resistance and the resistance of the field coils. Now that we are ready to interpret the above formula, let us simplify it by writing

$$k_1 = \sqrt{\frac{k_T}{k_E^2 k_\Phi}} \qquad \text{and} \qquad k_2 = \frac{R_a}{k_E k_\Phi}$$

so that

$$\omega_m = k_1 \frac{V_B}{\sqrt{T}} - k_2 \qquad (8\text{-}10)$$

expresses the desired relationship between speed (ω_m) and torque (T).

This formula shows the interesting feature of infinite speed when torque is zero, and has the general form shown in Fig. 8-12, for two values of V_a. Regardless of the value of V_B, current and torque are related by the same curve, shown dashed in the figure. The zero torque condition is never actually realized because there is always some friction in the machine, which has been neglected here.

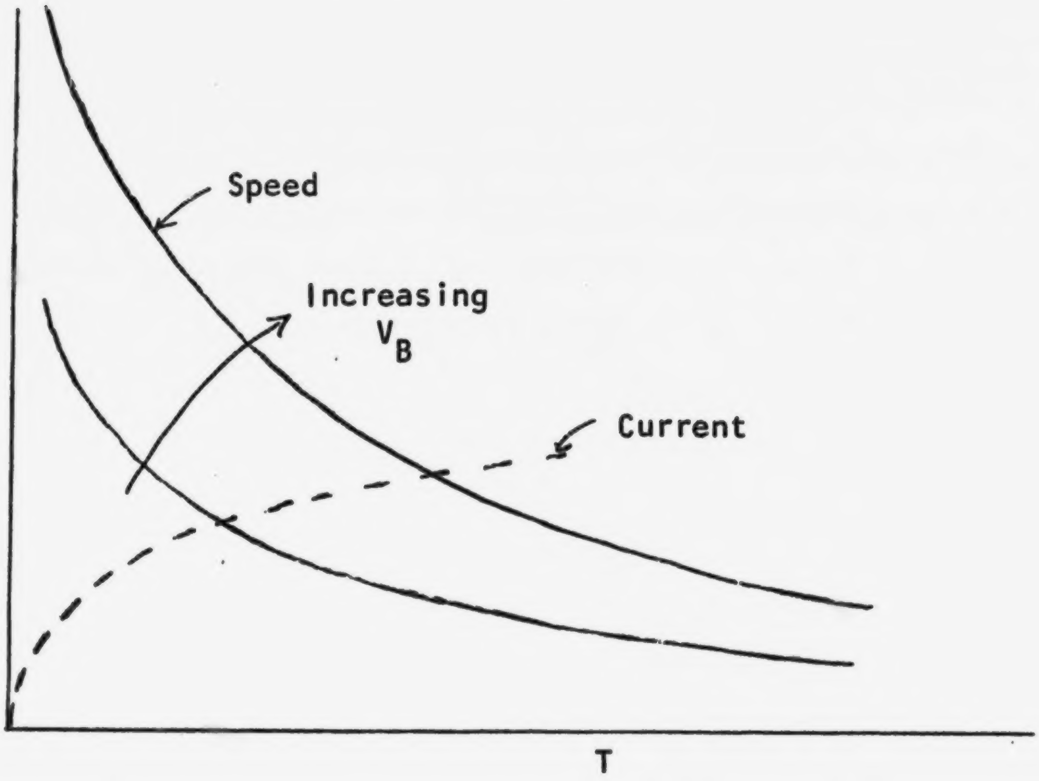


Figure 8-12.

Series connected motors inherently have more widely varying speed, as a function of torque, than shunt motors. Accordingly, they are useful in applications where load torque remains relatively constant over long periods of time, and where the load can never be removed. The latter condition is necessary to prevent "run away" at no load. Most small series motors will have enough losses due to friction to prevent damage if they are accidentally unloaded, but a large motor (1 hp. or more) can be very dangerous if allowed to run away, because centrifugal force can cause it to "explode". Railway traction motors are of the series type.

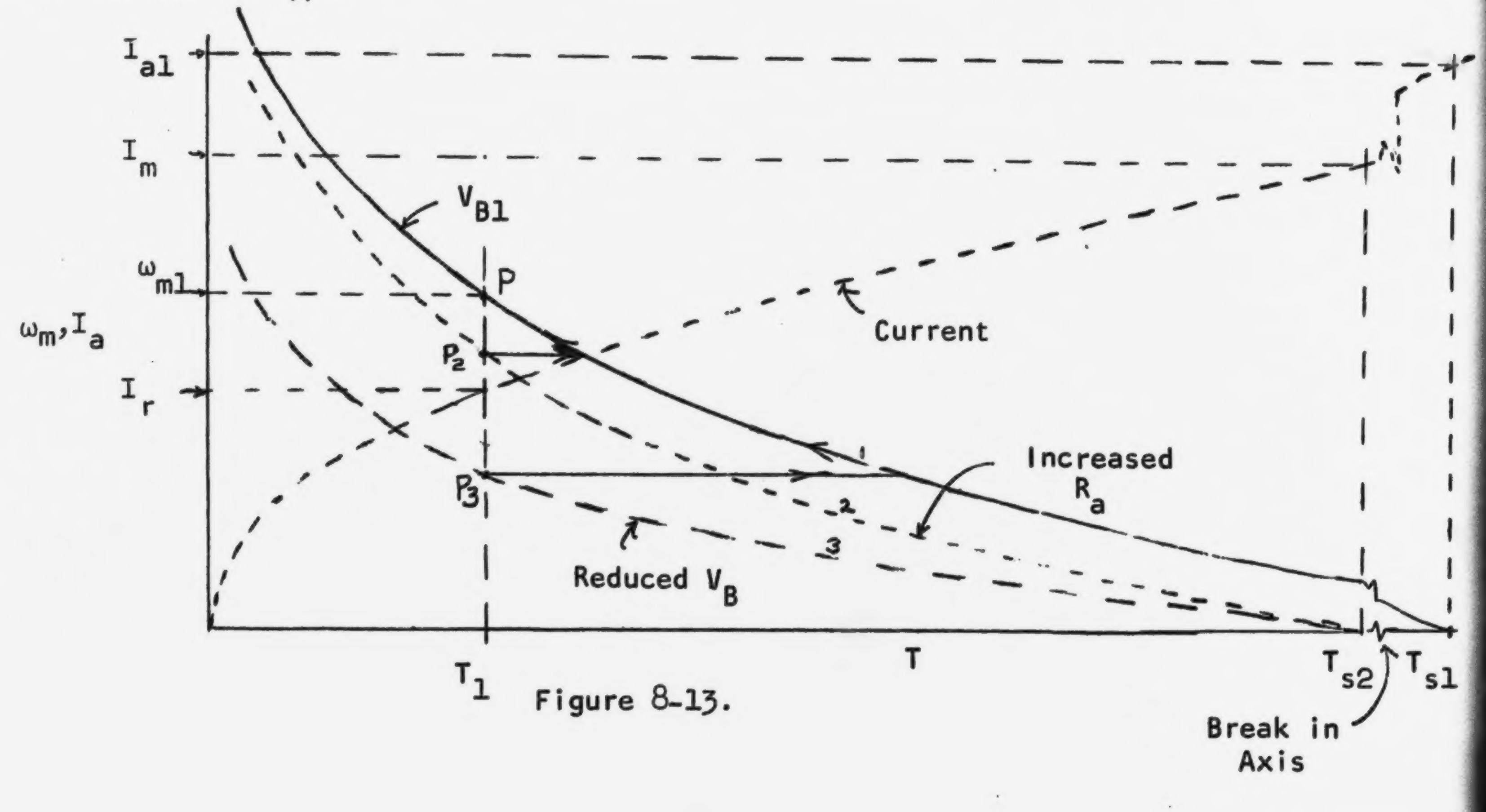


Figure 8-13.

Speed control, and starting conditions will be discussed in terms of Fig. 8-13. Let P represent a desired operating point for a load requiring a torque T_1 to be driven at a speed ω_{m1}. This will determine the particular value of source voltage V_{B1}, this being the voltage for the characteristic curve passing through point P.

The condition portrayed is the most severe possible in regard to starting, because the operating point is the full capacity of the machine, by which we mean that the armature current (as obtained from the I_a vs. T curve) is the maximum allowable value (I_r) for continuous running. As before, we shall assume that during the temporary starting interval the armature current can be allowed to reach a value, $I_m = 2I_r$.

If the motor were to be started "across the line", the starting torque would be T_{s1}, with a corresponding current I_{a1}, as indicated, which is in excess of I_m. The allowable starting torque T_{s2} can be obtained by projecting down from the current-torque curve at the point corresponding to I_m. Two curves, labeled (2) and (3) are shown which yield T_{s2} at zero speed. Curve (2) is obtained by shifting curve (1) downward a uniform amount throughout its extent. Referring to Eq. (8-10) we see that this change is accomplished by increasing k_2 (or increasing R_a). Curve (3) is obtained by multiplying $\omega_m + k_2$ by a constant factor. This explains why the reduction in going from curve (1) to curve (3) is greater at high values of ω_m. Referring to Eq. (8-10) again, it is seen that the change to curve (3) is obtained by reducing the source voltage V_B.

Thus, it is seen that there are two methods of holding the armature current within tolerable limits when the speed is zero, on starting. One method is to insert additional resistance in series with the motor, the other is to reduce the supply voltage.* Although curves (2) and (3) are slightly different, either one will provide adequate starting. The motor will be allowed to come up to constant speed (at points P_2 or P_3), and then the additional resistance will be taken out, or full voltage will be applied, at which time progress to the final operating point would be as indicated by the arrows.

---

*You may wonder why reduction of supply voltage was not mentioned in connection with starting a shunt motor. This would be possible if two sources were available, a fixed voltage for the field and a variable one to supply armature current. Reducing the voltage to both would result in weakening the field and perhaps would reduce the torque to the point where the machine would not start.

We now make an important observation. Since torque is proportional to the square of current, the starting torque is $T_{s2} = 4T_1$, and hence the torque available for acceleration is $4T_1 - T_1 = 3T_1$. This is the worst possible condition, where T_1 corresponds to rated current, and is the same case considered for the shunt motor, where it was found that the torque available for acceleration was T_1. Thus, the series motor supplies three times as much accelerating torque as the shunt motor, when started under conditions of constant load torque of maximum allowable value. For lighter starting load torque, the advantage of the series motor is slightly greater. The difference is, of course, due to the difference between the linear (T proportional to I_a) relation for the shunt motor and square law (T proportional to I_a^2) for the series motor.*

From the standpoint of getting the motor rotating, it makes little difference whether series resistance is inserted, on the supply voltage is reduced. Conditions affecting the choice can be appreciated when we consider speed control. It is evident from Fig. 8-13 that curves (2) and (3) both provide reduced speed for a given load, and therefore that the two methods of starting also comprise two methods of speed control.

A choice depends on considerations of efficiency. A series resistance absorbs power, and so variation of supply voltage is a more efficient way to control speed than by inserting a variable resistance. Thus, in applications where speed control is needed continuously, the increased efficiency of the variable voltage method may be definitive in indicating a choice of that method, although it involves a more expensive installation in the form of a separate motor generator set.**

---

*It is to be observed that the numbers computed above depend upon the arbitrary statement that starting current could be allowed to be twice the maximum allowable running current. If this factor had been three, the accelerating torque would be $3T_1-T_1=2T_1$, and $9T_1-T_1=8T_1$ respectively for the shunt and series cases.

**Or possibly a rectifier with controllable output voltage, if the d-c supply is obtained by rectifying an a-c source.

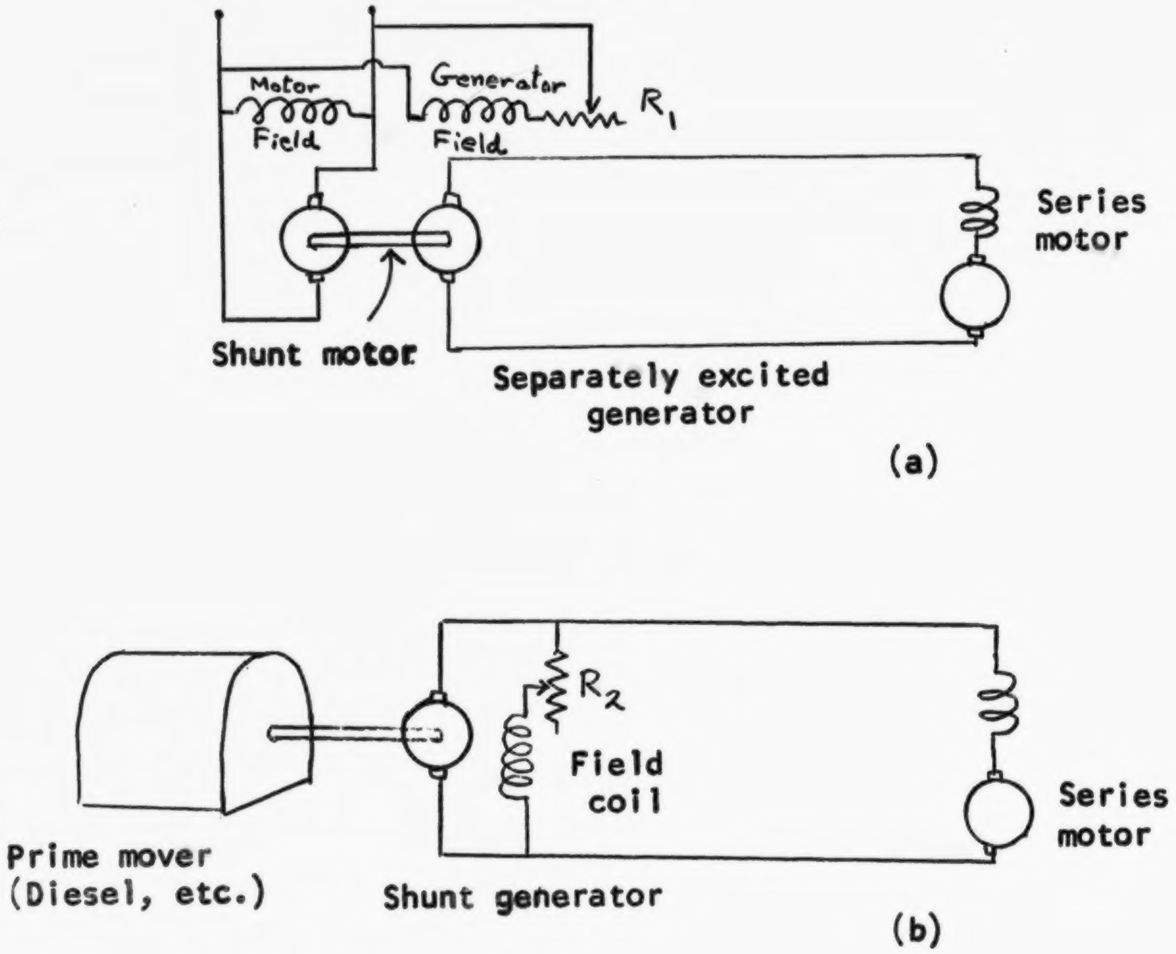

Motor Field
Generator Field
R_1
Series motor
Shunt motor
Separately excited generator
(a)
R_2
Field coil
Series motor
Prime mover (Diesel, etc.)
Shunt generator
(b)

Figure 8-14.

Two examples of variable voltage speed control are shown in Fig. 8-14. The system at (a) is used where d-c power is available from a constant voltage source. The source is connected to a nearly constant-speed shunt motor which drives a separately excited generator (meaning that the field current is independent of its own armature voltage). The strength of the field of this generator is controlled by a field rheostat R_1 which, therefore, also controls the voltage at the terminals of the generator armature. This is also the supply voltage for the series motor. A system like this would be applicable, for example, to the drive for a building elevator. The continuous use of such a system would warrant the initial expense of the motor-generator set which would not be needed if resistance control were used.

The system shown in Fig. 8-14b embodies the rudiments of a diesel-electric locomotive. In this case the generator voltage is determined by appropriate setting of the generator speed (as controlled by the engine throttle) and the setting of a field rheostat R_2.

8-6 Commutator a-c Motors

In the shunt and series d-c motors described in the previous sections it can be seen that the direction of rotation is independent of the polarity of the source. A change in source polarity will change the field flux direction and also the armature current direction. Reference to Fig. 8-4 will show that when both of these are reversed, the torque will still be in the same direction.

One might therefore expect that such motors would also operate satisfactorily on alternating current. This expectation is correct for a series motor, but not for a shunt motor. Let us first briefly consider why the shunt connection will not operate on a-c. The field coil and armature circuit, which are in parallel across the supply voltage, will each be inductive and carry currents which will not be in phase with the supply voltage, nor necessarily in phase with each other. On the other hand, in order to produce torque, it is necessary that the field current (and hence flux) and armature current shall be in phase (so that the product ΦI_a shall always be positive). Although this might be arranged in the shunt connection for a certain operating condition, any circuit change for the purpose of speed control or starting would involve changing armature or field resistances, and would disturb the phase relationship between flux and current. Furthermore, since the field coil has a high inductance, both armature and field currents would lag the voltage by nearly 90°. The result would be operation at low power factor, which means an excessively large current for the amount of power delivered. Thus, for these practical reasons, a shunt connection is not satisfactory for a-c operation.

The series motor does not suffer from these difficulties for the reasons that the series connection ensures that the field and armature current will be in phase, and the relatively few turns of a series field coil yields a low enough inductance so the power factor will not be excessively low.

This discussion is not meant to imply that any d-c series motor can be operated on a-c. The main difference is that for a-c operation the entire magnetic circuit must be laminated to reduce eddy current losses associated with the continual reversal of the magnetic field. In a d-c machine, the flux reverses in the rotor only, and so it must be laminated, but the stationary part of the magnetic circuit (yoke end pole pieces) can be solid iron. Thus, an a-c series motor can be operated on d-c, but not vice-versa.

Commutator a-c motors are particularly common in small sizes, being the typical motors for low power household applicances like vacuum cleaners, mixers, sewing machines, and the like.

It should be pointed out that a-c commutator motors are an exception to the general statement made in the introduction that there is no difference between a motor and a generator. If mechanical power is applied to drive an a-c commutator motor, it will produce d-c.

8-7 The Induction Motor

Applications requiring nearly constant speed a-c motors (similar to the d-c shunt motor) are served by a-c induction motors, which we shall now consider briefly. Unfortunately, an induction motor is much more difficult to analyze than its d-c counterpart, but at least enough theory can be presented here to discuss its important characteristics.

As a first step, consider Fig. 8-15a, which is a reproduction of Fig. 41 of the programmed text, with the slight modification of showing pairs of coil sides placed in slots. Figure 15b shows the same arrangement, uncluttered by the end connections. These figures represent the stator winding of an induction motor. Again referring to Fig. 15a, assume an a-c source is connected to terminals 1 and 1', and let i_1 have the reference direction shown. An alternating current will flow in the individual coil sides shown in Fig. 15b, with reference directions as indicated by the X and · symbols.

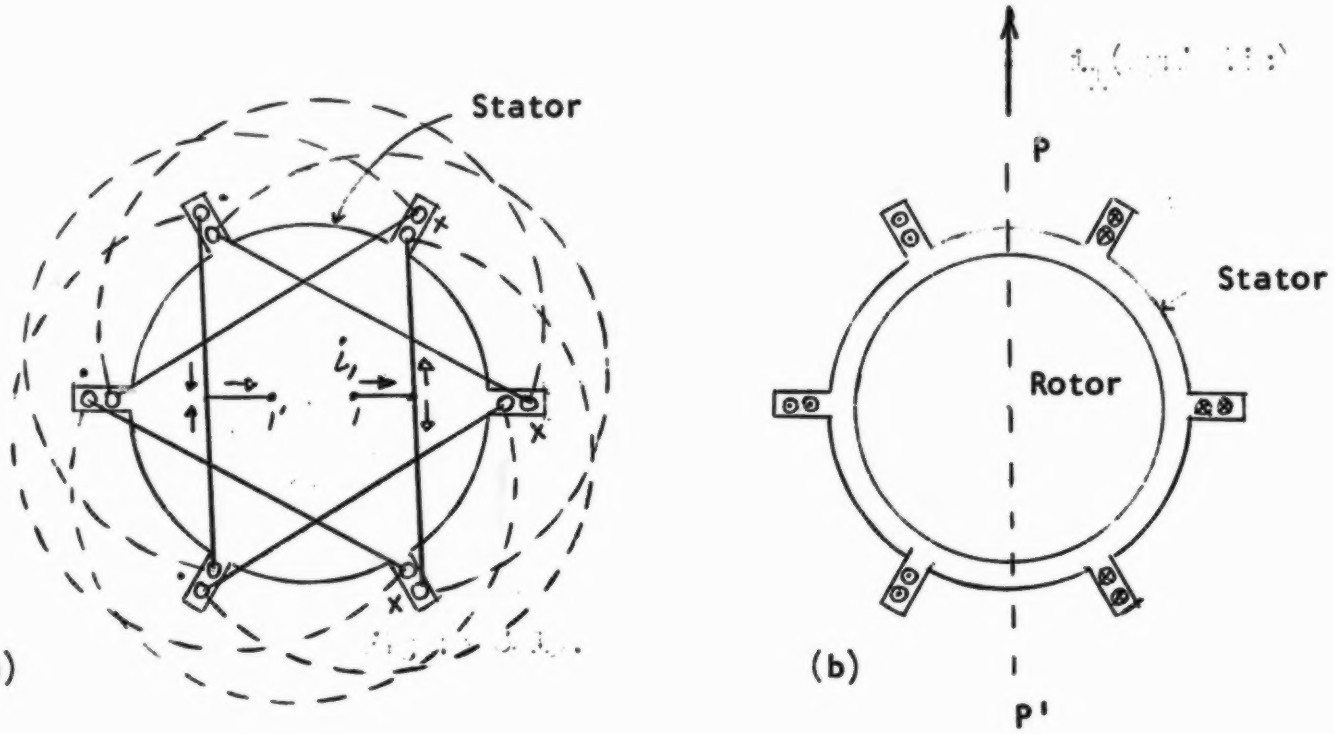


Figure 8-15.

This figure serves to remind you that coil connections are such that at any instant of time currents to the right and left of line P-P' always flow in opposite directions, and that the whole winding acts like a coil having P-P' as an axis and producing flux having a reference direction indicated by the arrow.

Twenty or more coils would be a more reasonable number for a practical machine. Figure 8-16a shows the coil sides for a ten coil winding. This figure is introduced to develop the notion that in an actual machine coil sides are separated by a small angle δ, and therefore that the actual condition is not much different from the idealized continuous distribution of current shown in Fig. 8-16b. This idealization is introduced because it makes the subsequent analysis simpler than it would be if we were to bother with the details of individual coils. In other words, we are proceeding on the basis of the hypothesis that this approximation will lead to an adequate theory, with justification depending upon agreement with experimental observation.

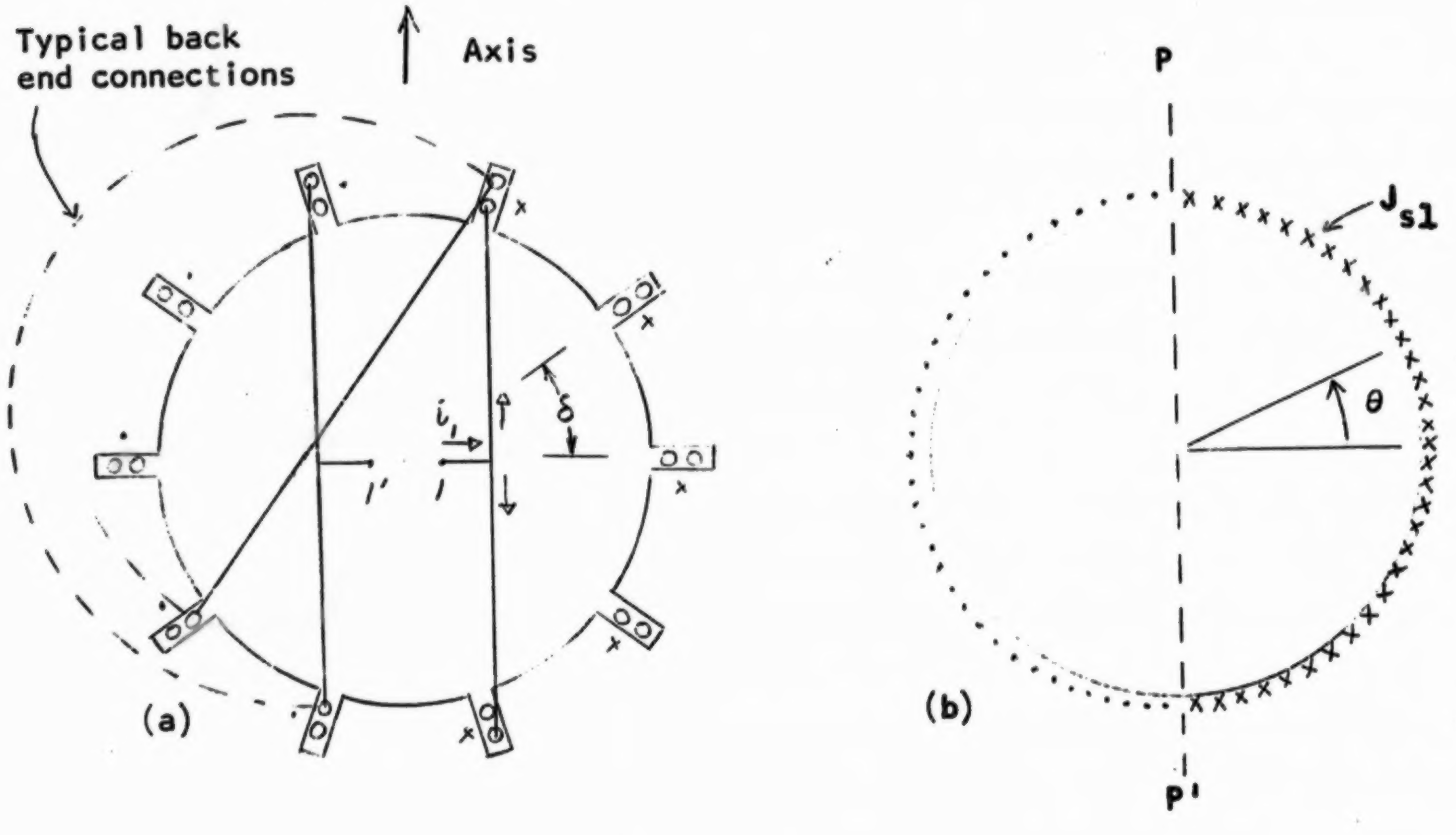


Figure 8-16.

The symbol shown for current in Fig. 8-16b is J_{s1}. This is a surface current density, measured in amperes per meter of circumference. To explain what we mean by this, suppose the winding of Fig. 8-16a has N coils and C turns per coil. There will be NC conductors on one side of line P-P', and each conductor carries a current $i_1/2$ (Why ?). Thus, the total current on one side will be $NCi_1/2$. For Fig. 8-16b to be equivalent, it must have the same total current on each side of line P-P'. This is $\pi r J_{s1}$, and hence we see explicitly how to find J_{s1}; namely,

$$J_{s1} = \frac{NCi_1}{2\pi r} \tag{8-11}$$

Having shown how machine parameters determine J_{s1}, we shall use J_{s1} as the prime quantity from which to start the analysis. Meanwhile, we must not forget that i_1 is an alternating current, and therefore that J_{s1} is also a sinusoidal quantity. It is also a function of θ, and can be written

$$J_{s1} = (J_m \cos \omega_e t) f(\theta) \tag{8-12}$$

where the $J_m \cos \omega_e t$ factor accounts for the time variation and $f(\theta)$ describes the variation with θ. For Fig. 8-16b, $f(\theta)$ is the square wave shown by the solid line in Fig. 8-17.

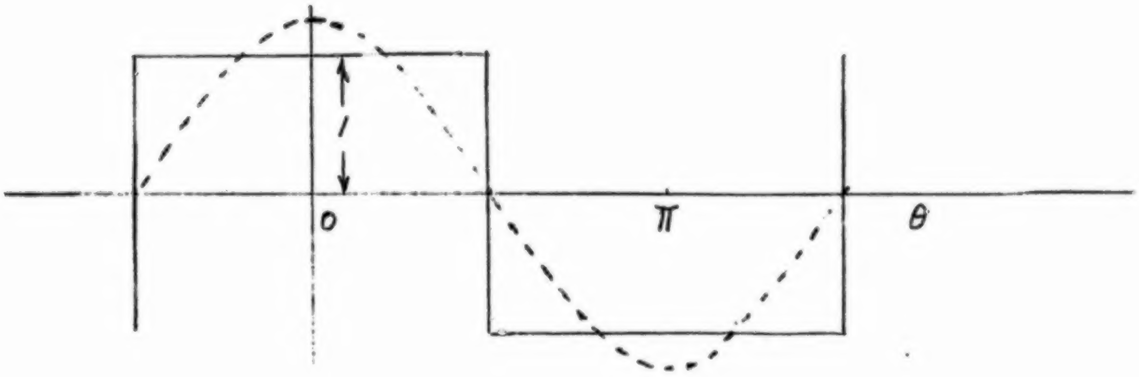


Figure 8-17.

At this point we make a further approximation, by using the dotted cosine function shown in the figure, to approximate the square wave.* Thus, we shall

---

*This approximation is not necessary, but permits a simplified treatment in this text which does not presuppose a knowledge of Fourier Series. A more complete analysis would retain the square wave which can be represented by the Fourier series

$$\frac{4}{\pi}(\cos \theta - \frac{1}{3} \cos 3\theta + \frac{1}{5} \cos 5\theta + \ . \ . \ .)$$

Thus it is seen that the subsequent treatment will use only the first term of this series.

use

$$J_{s1} = \frac{4J_m}{\pi} \cos \omega_e t \cos \theta \qquad (8\text{-}13)$$

and, in view of Eq. (8-11), it is also given by

$$J_{s1} = \frac{NCI_m}{\pi^2 r} \cos \omega_e t \cos \theta \qquad (8\text{-}14)$$

where I_m is the maximum input current.

The next step is to use the identity cos x cos y = (1/2)[cos (x-y) + cos(x+y)] to write Eq. (8-13) as

$$J_{s1} = \frac{2J_m}{\pi}\left[\cos(\omega_e t-\theta) + \cos(\omega_e t+\theta)\right] \qquad (8\text{-}15)$$

This is an important step because the dependence of J_{s1} on t and θ has been converted from a product of a function of t by a function of θ to a sum of two functions in which t and θ are combined in the arguments $\omega t - \theta$ and $\omega t + \theta$.

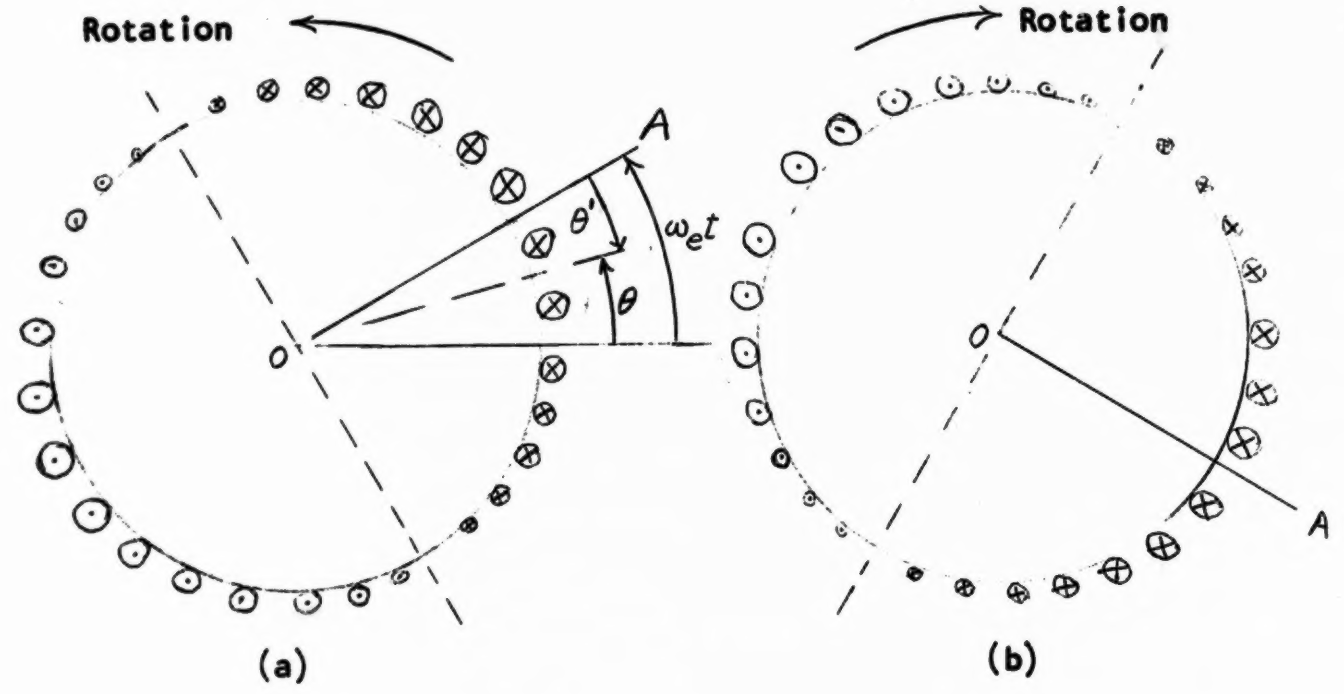
We shall interpret the term $\cos(\omega_e t-\theta)$ in terms of Fig. 8-18a which portrays a fictitious cylindrical distribution of constant currents which is rotating at angular velocity ω_e. The position of this rotating system of currents is defined by the position of its axis is of symmetry 0-A. In time t this axis has rotated through an angle $\omega_e t$, as indicated. With respect to angle θ' measured from an axis fixed in the rotating cylinder, J_{s1} is a function of θ' only, namely

$$J_{s1} = \frac{2J_m}{\pi} \cos \theta'$$

This relationship is suggested by the varying sizes of the circles which represent currents. At a fixed observation point, at angle θ on the stator, we see from the diagram that $\theta' = \omega_e t-\theta$, and so the current density observed will be

$$\frac{2J_m}{\pi} \cos \theta' = \frac{2J_m}{\pi} \cos(\omega_e t-\theta)$$

which is the first term of Eq. (8-15). Accordingly, that part of the variation of J_s due to $\cos(\omega_e t-\theta)$ is accurately portrayed by the conceptual picture of a rotating system of currents which are constant in time. The time variation is provided by the rotation. Similarly, the $\cos(\omega_e t+\theta)$ term can be viewed as due to a similar set of currents rotating in the opposite direction, as suggested by Fig. 8-18b.

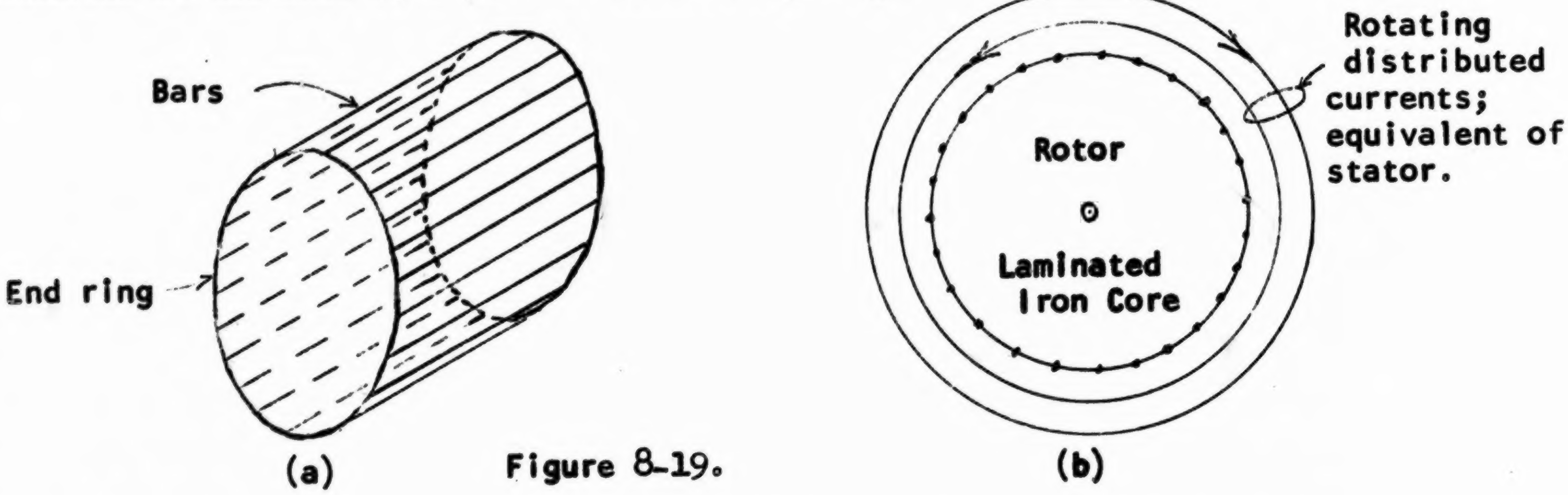

Rotation
A
θ'
$\omega_e t$
θ
O
Rotation
O
A

(a) (b)

Figure 8-18.

We have now reached the important conclusion that a sinusoidally time varying current distribution in a fixed winding (as represented by Fig. 8-16b) can be replaced conceptually by the superposition of two oppositely rotating constant distributions, as in Fig. 8-18.

This analysis permits us to give an intuitive description of how an induction motor operates. A rotor carrying a winding short-circuited upon itself is placed inside the stator we have just described. The winding may consist of coils placed in slots, like the stator winding, or it may be a "squirrel cage" winding as portrayed in perspective view in Fig. 8-19a, in which a series of equally spaced bars are placed in rotor slots, and are connected at the ends by a pair of conducting rings.


Bars
End ring
Rotating distributed currents; equivalent of stator.
Rotor
Laminated Iron Core

(a) Figure 8-19. (b)

Figure 8-19b describes the state of affairs so far established. The stator is equivalent to two oppositely rotating current distributions, as if there were two oppositely rotating systems of coils. In view of the opposing directions, the rotor will not be affected when stationary, but if it is started in either direction, rotor currents will be induced which will react strongly with the current distribution rotating in that direction, and weakly with the current distribution rotating in the opposite direction. This will result in a net torque in the direction of rotation, which will keep it going. Such a motor will run in whatever direction it is started, but will not be self starting.

The above facts are crucial, because they indicate what must be done in order to make an induction motor that will be self starting and will run in only one direction. It is necessary to eliminate one of the counter-rotating current distributions. There is more than one way to do this, one of them being illustrated by Fig. 8-20, which shows a reproduction of Fig. 8-16a in solid lines, and a second equivalent set of coil slots drawn dotted. The second set of coils are shifted 90° in a counter clockwise direction in space, and carries a current which lags i_1 by 90°. That is,

$$i_2 = I_m \cos(\omega_e t - \tfrac{\pi}{2})$$

This can be approximately described by a second equivalent continuous current distribution

$$\begin{aligned} J_{s2} &= \frac{4J_m}{\pi} \cos(\omega_e t - \tfrac{\pi}{2}) \cos(\theta - \tfrac{\pi}{2}) \\ &= \frac{2J_m}{\pi}\left[\cos(\omega_e t - \theta) + \cos(\omega_e t + \theta - \pi)\right] \end{aligned} \qquad (8\text{-}18)$$

However, $\cos(\omega t + \theta - \pi) = -\cos(\omega t + \theta)$, and so the above reduces to

$$J_{s2} = \frac{2J_m}{\pi}\left[\cos(\omega_e t - \theta) - \cos(\omega_e t + \theta)\right] \qquad (8\text{-}19)$$

If both windings are simultaneously energized, the total equivalent current distribution is

$$\begin{aligned} J_s &= J_{s1} + J_{s2} \\ &= \frac{4J_m}{\pi} \cos(\omega_e t - \theta) \end{aligned} \qquad (8\text{-}20)$$

which also represents a fixed distribution rotating in only one direction (counter-clockwise).

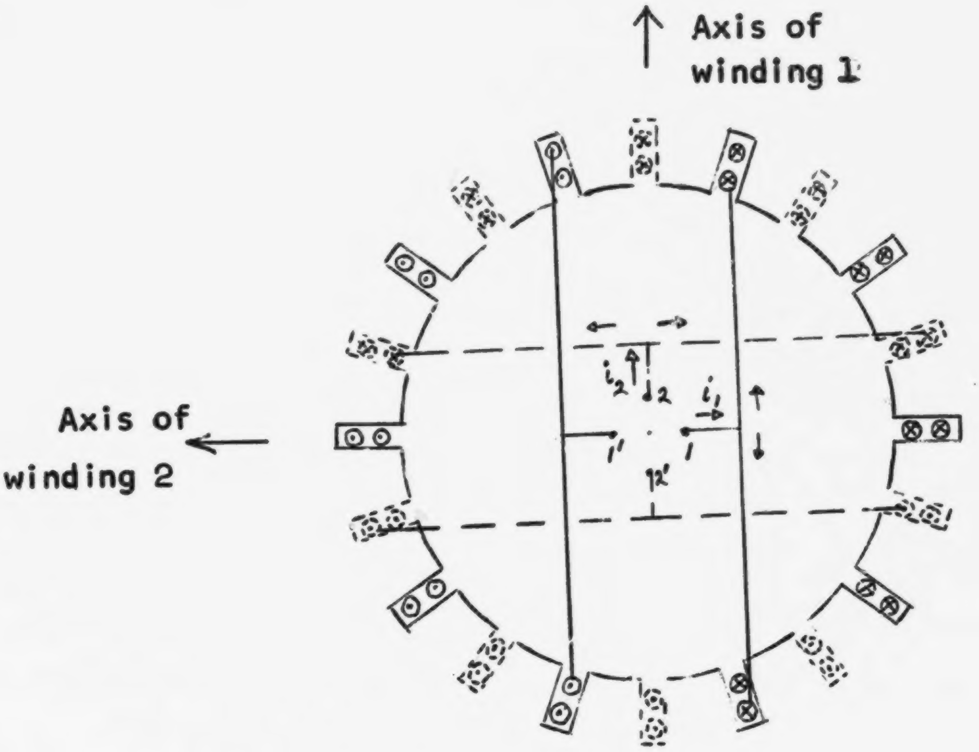


Figure 8-20.

A similar analysis will show that if current i_2 were to lead i_1 by 90°, rotation would be in the clockwise direction. This fact indicates how such a motor can be reversed. A change from lead to lag in winding (2), compared to winding (1), can be accomplished merely by interchanging its two terminal leads.

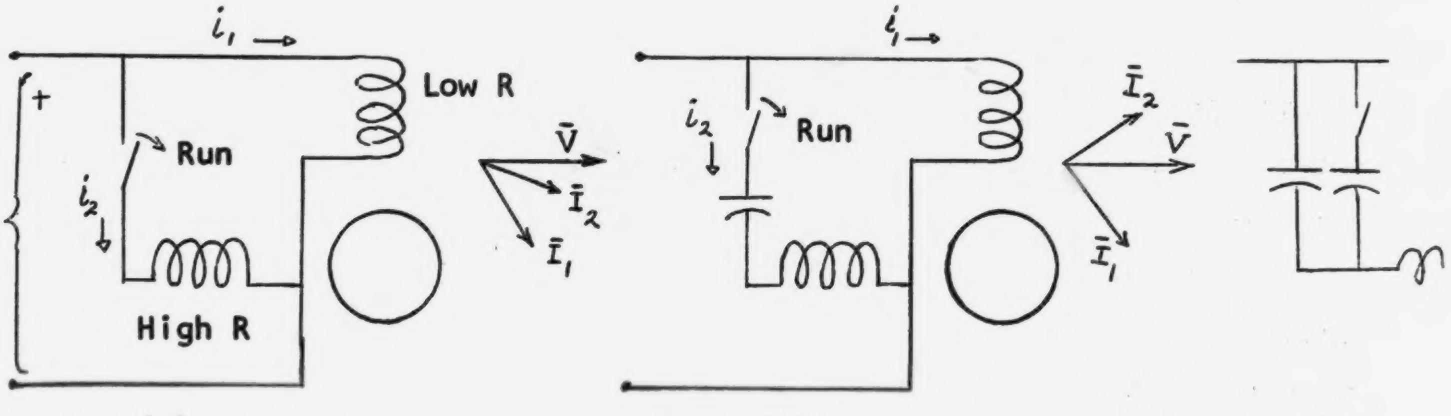
Such a motor is called a two-phase motor, because it has two separate windings supplied by two separate sources. In practice further modificiations are made, as described below.

8-8 Single-phase motors

In small and medium sizes (usually a fraction of a horsepower) it is convenient to have motors that do not require the complexity of two separate power supplies. It will be recalled that a single winding motor (called a single-phase motor) will run once it is started. Such a motor can be made self starting by adding a second winding, which is used only during the starting period. This raises the question of how to arrange for the 90° phase difference in currents i_1 and i_2, when supplying them from the same source.

There are two practical solutions to this problem. One is to make the second winding of high resistance, by constructing it of smaller wire, and to connect the two windings in parallel across the source terminals, as shown in Fig. 8-21a. Current i_1 flows in a circuit of high L and low

R, and will lag the voltage by nearly 90°. On the other hand, winding (1) is of high resistance and so its current will be nearly in phase with the voltage. Thus, i_2 will lead i_1 by something less than 90° (perhaps 60°), as suggested by the phasor diagram in Fig. 8-21a, and so the condition required by the theory will be approximated, to a sufficient degree that the motor will start if the required starting torque is not too great.* After the motor has started, a centrifugal switch opens the starting winding circuit. Continuous operation with the two windings would not be efficient because of the high resistance of the starting winding, and also because as the motor speeds up winding (1) becomes less inductive and the phase difference would become much less than 90°.



(a) (b) (c)

Figure 8-21.

The above describes the original design for single phase motors (1/8 to 1/2 hp.) until the advent of electrolytic capacitors which made it possible to use a capacitor in series with the second winding, as shown in Fig. 8-20b. By suitable choice of the capacitor size, current i_2 can be made to lead i_1 by exactly 90° at the start, since with a sufficiently large capacitor i_2 can actually lead the voltage, as shown in the phasor diagram of Fig. 8-21b.

---

*Observe that when we introduced the two-phase motor it was indicated that i_2 should lag i_1 by 90°, whereas this description shows it leading by nearly 90°. This is incidental, since it only affects the direction of rotation, and can be compensated for by a simple interchange of connections to winding (2).

Under running conditions, the winding (2) circuit can be opened by a centrifugal switch, as in the previous case. However, improved performance can be obtained by retaining the second winding under running conditions. In that case, under running conditions, i_1 will be more nearly in-phase with the voltage which means that i_2 must lead the voltage by a larger angle on "run" than on "start". This change can be accomplished by having a smaller series capacitor for the "run" condition. This can be accomplished as shown at (c) in Fig. 8-21. Two capacitors are in parallel on "start", and then as the motor comes up to speed a centrifugal switch opens, removing one of them from the circuit. The capacitor-start capacitor-run motor just described is a popular type of motor for applications such as domestic refrigerators. It has the advantage of providing more torque for a given size, because the crossed field arrangement of a two-phase motor is retained. It has the further advantage of operating at nearly unity power factor, by virtue of the fact that the current in one winding leads the voltage and therefore can cancel the lagging component of current in the other winding.

Both types are called split-phase motors. They operate from a single-phase source, but this source is "split" to yield effectively two phases within the motor.

A third type of single-phase induction motor is worth mentioning because it is quite popular in small sizes (small fan motors, phonograph turntable drives, etc.). Such a motor does not have a distributed winding, and has definite pole pieces, much as in the d-c case. However, as indicated in Fig. 8-22, part of each pole is encircled by a heavy copper band. This is called a shading coil. As the a-c flux changes in that part of the pole encircled by

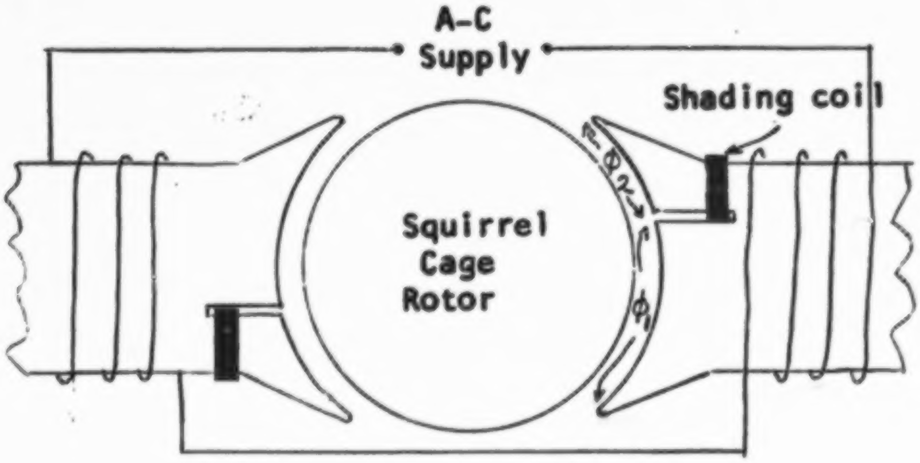


Figure 8-22.

the shading coil, it induces a current in the shading coil which, by virtue of Lenz's law, tends to oppose this change. The result is that flux ϕ_2 will lag flux ϕ_1. It is also displaced in space, and hence the combination produces an approximation of the condition of a set of rotating windings. There are high losses in the shading coils, which remain continuously operative. Thus, such a motor is inefficient, but its great simplicity makes it attractive for applications where the power requirement is low and where, therefore, a low efficiency is not serious.

8-9 Three phase motors

The split-phase and shaded-pole motors described above generally are inefficient and are larger in size, compared with motors supplied from a two-phase source. Therefore, in sizes larger than about one horsepower, it is necessary to have a method of supplying power directly to the second winding. We shall now discuss how this is done in practice. The two coil arrangement of Fig. 8-20 is of tutorial value, as we have used it above, and also is of historical interest, having been a practical solution in some early motors.

In practice, the second winding of Fig. 8-20 is replaced by two windings, giving a total of three windings and creating what is called a three-phase machine.

Figure 8-23 symbolically represents the winding arrangement. There are three separate windings, labeled (1), (2), and (3), which are insulated from each other. Winding (1), for which circles are used to represent coil sides in slots, is the same as in Fig. 8-16a. Winding (2), represented by squares, has its axis displaced 120° from the axis of winding (1). As you study this figure, observe that every third position belongs to winding (1), and that likewise the coil sides of winding (2) occupy every third position. Also, observe the locations of the X and dot symbols, in relation to the current reference direction at the one terminal shown for each winding. The connections to the second terminals are omitted, although parts of the end connections are shown as dotted lines. (Refer to Fig. 8-16a to refresh your memory as to where the second terminal is connected.) Finally, winding (3) is similar, but has its axis shifted another 120°. The symbolism whereby circles, squares, and triangles represent coil sides is indicated by the insert.

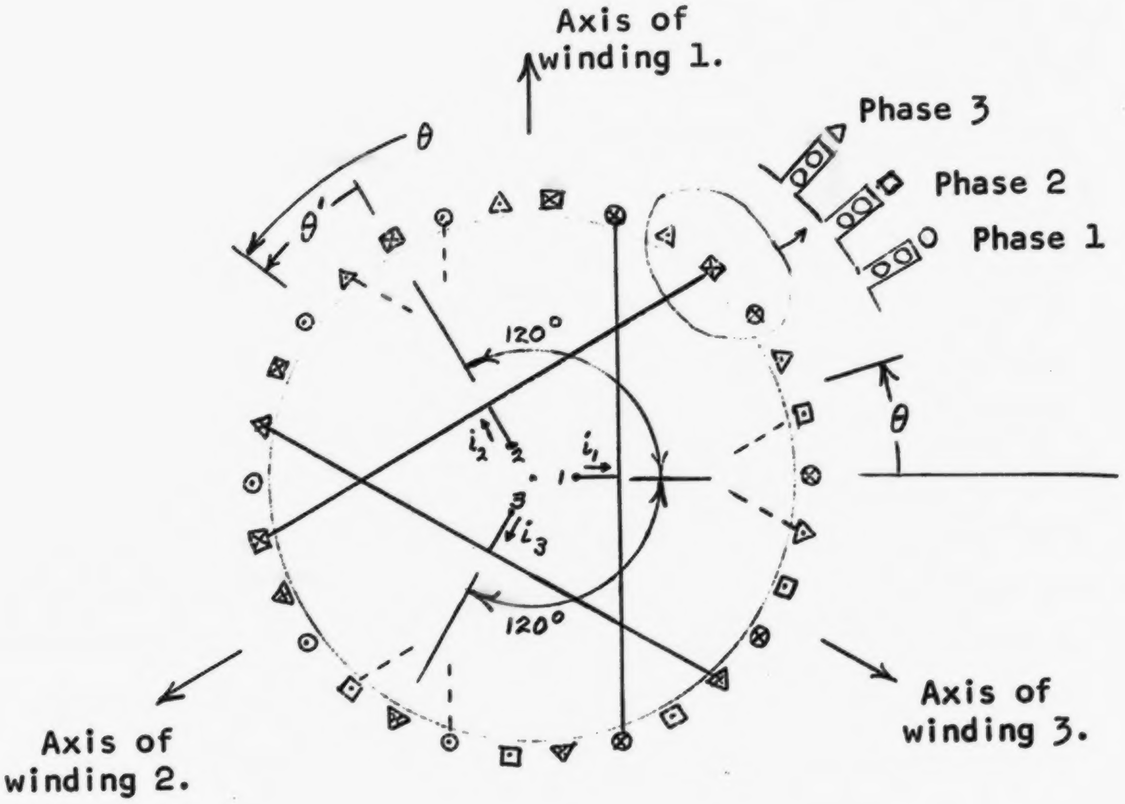

Axis of winding 1.
Phase 3
Phase 2
Phase 1
θ
θ'
120°
i_2
2
i_1
1
3
i_3
120°
θ
Axis of winding 3.
Axis of winding 2.

Figure 8-23.

To begin the analysis, you are reminded that winding (1) produces an equivalent current distribution

$$J_{s1} = \frac{2J_m}{\pi}\left[\cos(\omega_e t-\theta) + \cos(\omega_e t+\theta)\right]$$

and that winding (2) of Fig. 8-20 introduces a second distribution

$$J_{s2} = \frac{2J_m}{\pi}\left[\cos(\omega_e t-\theta) - \cos(\omega_e t+\theta)\right]$$

such that the negative sign of the second term cancels the unwanted second term of the expression for J_{s1}. We shall now show how Fig. 8-23 accomplishes a similar cancellation, if the sources to the three windings are adjusted to have equal amplitudes, and phase relations in accordance with the expressions

$$i_1 = I_m \cos \omega_e t \qquad \text{(a)}$$

$$i_2 = I_m \cos(\omega_e t - \tfrac{2\pi}{3}) \qquad \text{(b)} \qquad \text{(8-21)}$$

$$i_3 = I_m \cos(\omega_e t + \tfrac{2\pi}{3}) \qquad \text{(c)}$$

Each of these will be represented by an approximately equivalent distribution, to be represented by J_{s2} and J_{s3}. (Observe that the subscript 2 now refers to Fig. 8-23, rather than Fig. 8-20.) The equivalences are in accordance with the treatment of Fig. 8-16, the only differences being in angular orientations of the windings in space, and the phase positions of the currents. For winding (2) the current distribution is

$$J_{s2} = \frac{4J_m}{\pi} \cos(\omega_e t - \frac{2\pi}{3}) \cos(\theta - \frac{2\pi}{3}) \tag{8-22}$$

The reason for the $(-2\pi/3)$ term in the time varying term is evident, since it is the phase angle established in Eq. (8-21b). In regard to $\cos(\theta - 2\pi/3)$, observe from the figure that winding 2 is oriented in such a way that its distribution is $\cos(\theta')$. However, it is evident from Fig. 8-23 that $\theta' = \theta - 120^\circ$, or $\theta - 2\pi/3$ in radian measure. The identity $\cos x \cos y = (1/2)\left[\cos(x-y) + \cos(x+y)\right]$ is used on Eq. (8-22) to give

$$J_{s2} = \frac{2J_m}{\pi} \left[\cos(\omega_e t - \theta) + \cos(\omega_e t + \theta - \frac{4\pi}{3})\right]$$

$$= \frac{2J_m}{\pi} \left[\cos(\omega_e t - \theta) - \frac{1}{2}\cos(\omega_e t + \theta) - \frac{\sqrt{3}}{2}\sin(\omega_e t + \theta)\right] \tag{8-23}$$

where the factors $(-1/2)$ and $(-\sqrt{3}/2)$ are the cosine and sine of $4\pi/3$, respectively.

The third winding is treated in an exactly similar fashion, to yield a distribution

$$J_{s3} = \frac{4J_m}{\pi} \cos(\omega_e t + \frac{2\pi}{3}) \cos(\theta + \frac{2\pi}{3})$$

$$= \frac{2J_m}{\pi} \left[\cos(\omega_e t - \theta) + \cos(\omega_e t + \theta + \frac{4\pi}{3})\right]$$

$$= \frac{2J_m}{\pi} \left[\cos(\omega_e t - \theta) - \frac{1}{2}\cos(\omega_e t + \theta) + \frac{\sqrt{3}}{2}\sin(\omega_e t + \theta)\right] \tag{8-24}$$

The sum of Eqs. (8-23) and (8-24) is

$$J_{s2} + J_{s3} = \frac{2J_m}{\pi}\left[2\cos(\omega_e t-\theta) - \cos(\omega_e t+\theta)\right]$$

and when this is combined with J_1, the $\cos(\omega_e t+\theta)$ term cancels to give

$$J_s = J_{s1} + J_{s2} + J_{s3} = \frac{6J_m}{\pi}\cos(\omega_e t-\theta) \tag{8-25}$$

Thus it has been shown trigonometrically that when all three windings act at once, the result is a counter-clockwise rotating equivalent current distribution like Fig. 8-18a. The direction of rotation resulted from our choice of phase relationships for the currents. If we had used $i_2 = I_m \cos(\omega_e t+2\pi/3)$ and $i_3 = I_m \cos(\omega_e t-2\pi/3)$, the $\cos(\omega_e t-\theta)$ terms would have cancelled, leaving $\cos(\omega_e t+\theta)$ in Eq. (8-24). This would represent a rotation in the clockwise direction.

Equation (8-25) for the three-phase machine, is to be compared with Eq. (8-20) for the two-phase case. It is evident that the only change is that the factor 4 has been replaced by 6. The comparatively simple results expressed by Eqs. (8-20) and (8-25) are important because they show that either the two-phase or the three-phase windings, when properly exicted, are equivalent to a current distribution rotating in one direction only. Thus, when a rotor is placed under the magnetic influence of such a set of windings it will experience a starting torque, and will rotate in only one direction, this direction being determined by the phase relationships among the stator currents.

A word is in order concerning how power is supplied to a three-phase motor. In Fig. 8-24a the three motor windings, which are represented symbolically, are connected by separate circuits to three generators rotating on the same shaft. The common generator shaft is necessary to ensure the desired 120° phase differences among the currents in the three circuits.

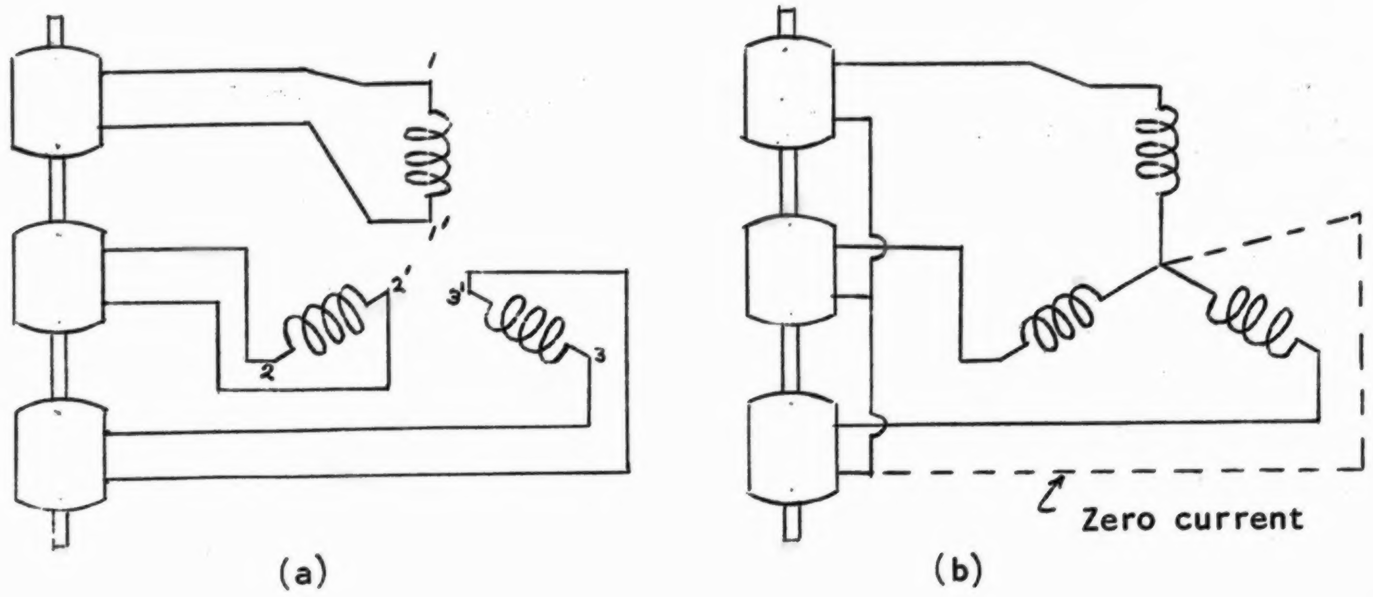


Figure 8-24.

All these wires are not necessary. To see why, observe that the bottom wires of each pair can be combined into a single wire, as in Fig. 8-24b. Each circuit can still operate independently, although the three circuits are now electrically connected on one side.* Furthermore, from Eqs. (8-21) we can write

$$i_1 = I_m \cos \omega_e t$$

$$i_2 = I_m \cos(\omega_e t - \frac{2\pi}{3}) = I_m \left[-\frac{1}{2}\cos \omega_e t + \frac{\sqrt{3}}{2}\sin \omega_e t\right]$$

$$i_3 = I_m \cos(\omega_e t + \frac{2\pi}{3}) = I_m \left[-\frac{1}{2}\cos \omega_e t - \frac{\sqrt{3}}{2}\sin \omega_e t\right]$$

and so it is seen that

$$i_1 + i_2 + i_3 = 0 \qquad (8\text{-}26)$$

Thus, no current will flow in the common (dashed) wire, and so it can be omitted.

We have arrived at the conclusion that power can be supplied to a three-phase motor through three wires. To round out this brief discussion of three-phase power transmission, it is to be mentioned that the three generators do not need to be physically separate as in Fig. 8-24. They can be incorporated on a single stator, and always are. In fact, the three-phase system of windings we have described for the induction motor will also serve as the set of three windings of a three phase generator.

---

*The circuits are independent if the impedance of the common wire is negligible.

The principles involved in the three-phase induction motor provide the explanation why all commerical power systems are three-phase. This is why the transmission lines you see over the countryside always have three wires (or multiples of three).

8-10 Induction Motor Speed-Torque Curves

It is beyond the scope of this treatment to derive the equation for the speed-torque characteristic of an induction motor. We therefore give typical examples without derivation, in Fig. 8-25. Curve A is typical for a three-phase motor with squirrel cage rotor. In the running range, speed is relatively constant. However, if the load torque is increased too far, the knee of the curve will be reached (this is called the pull-out torque) and the motor will suddenly come to a stop. This phenomenon is due to changing phase relationships in the rotor currents. Such a characteristic is not satisfactory for starting against a large torque. A squirrel cage induction motor would not be satisfactory in electric traction applications. They are satisfactory, however, in such an application as a lathe drive, where starting is only against the friction of the lathe, the cutting load being applied later.

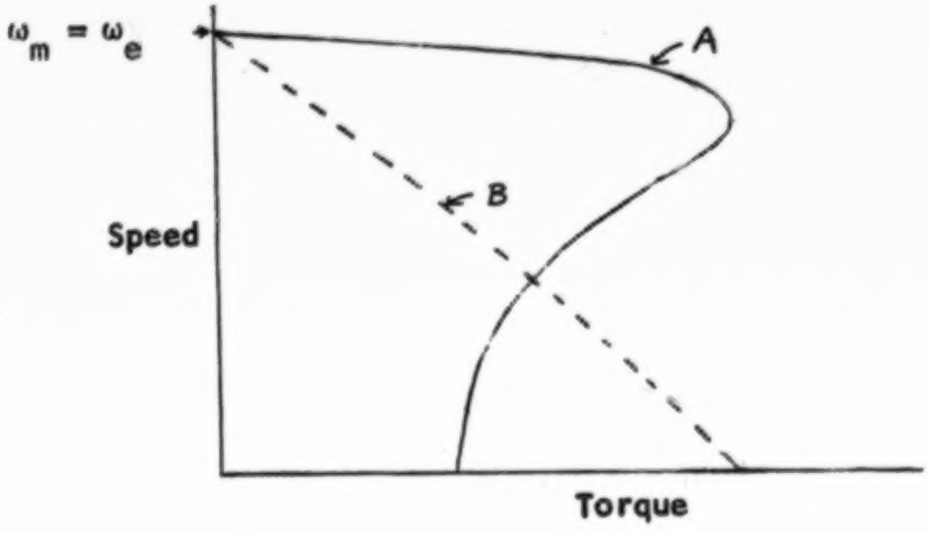


Figure 8-25.

The starting torque of an induction motor can be improved by increasing the electrical resistance of the rotor conductors. For this reason, some motors are made with a wound rotor (with a winding essentially like the stator) with connections to three external variable resistors through a set of slip rings. By increasing this resistance, a characteristic like B can be obtained, giving maximum starting torque. After the motor is running, the external resistance can be removed, causing the characteristic to become like curve A.

Concerning current and power, it is recalled that a-c power is given by VI cos θ, where θ is the phase difference between voltage and current. If V and I are for any one phase, the total power is

$$P = 3VI \cos \theta$$

The current of an induction motor does not vary widely from no load to full load, as in the case of d-c machines. The no-load current is large, of the order of half the full-load current. Most of the change in power comes about through change in phase angle. At no load (running but with no load torque) the current will lag the voltage by nearly 90°, and at full load an angle of 20° is reasonable.

These relatively large phase angles are a disadvantage which cause considerable difficulty to power companies. It may be necessary to install capacitors in the vicinity of large industrial loads, to compensate for the low power factor caused by a large number of induction motors.

In regard to starting, the use of rotor circuit resistance in a wound rotor machine has been mentioned. Small squirrel cage motors (up to about 2 hp.) can be started at full line voltage. However, it may be necessary to remove protective devices (fuses or circuit breakers) in order to prevent their operating on the heavy current that flows for a few cycles. Larger squirrel cage motors are started by applying reduced voltage, through a bank of transformers.

8-11. Motor Ratings

In the discussion of starting d-c motors it was pointed out that starting current can be permitted to exceed rated current for operation at full load. The name-plate current rating of a machine is determined by the allowable temperature rise which, in turn is primarily influenced by the insulation used in the windings.

Rated current is based on continuous operation. Momentary loads can cause the current to exceed rated value by considerable amounts, if they endure only for a few seconds. In order to obtain good service from a motor, it is necessary that it operated within its name-plate current ratings.

1 Operation of a motor at other than its rated voltage is not generally
recommended, except in the case of d-c motors which can be operated at reduced
voltage. An induction motor should not be operated for long at reduced voltage.
This is because its speed is largely determined by the line frequency, and
2 therefore with a given load its power output is nearly independent of voltage.
If the voltage is reduced, the current must increase, to maintain the power,
and the maximum allowable current may be exceeded.

8-12. Number of Poles

3 All of the discussion in this text (and the programmed text) assumes
machines of two poles. However, four, six, etc. pole machines are possible.
The characteristics of d-c machines are independent of the number of poles.
The only things affected are the armature winding and voltage and current
4 ratings.

Although poles are not physically present in an induction motor, the
equivalent of a pole is defined by the extent of the circumference of the
stator subtended by one coil. In our illustrations, this has been 180°,
5 which defines a two-pole motor. When there are four poles, the angle
subtended is 90°, etc. In our example, the equivalent current distribution
rotates at an angular speed ω_e, and this is the maximum speed of the rotor.
With a four pole winding, the rotating current distribution will have two
6 cycles of current variation (variation with angle) instead of one cycle as
in the two pole case. Functions like

$$\cos(\omega_e t - 2\theta)$$

7 will appear. Such a wave moves with angular speed $\omega_e/2$. Likewise, with six
poles the speed is $\omega_e/3$, and in general is ω_e/(number of pole pairs).

The speeds discussed here are the maximum (or synchronous) speeds.
Actual speeds are usually higher than 95% of synchronous speed. At 60 cps.
8 $\omega_e = 377$ radians per second or 3600 RPM. This is approximately the speed
of a two-pole machine. Four and six pole machines are very popular, and
have approximate operating speeds, respectively, of 1800 and 1200 RPM.
Thus, if you see a name-plate indicating a speed of 1740 RPM, this is a
9 four-pole motor.

BASIC SEMICONDUCTOR THEORY

Each of the following paragraphs presents a few ideas that are essential to an understanding of semiconductor theory. Be sure you know the material in each paragraph before going on to the next.

1. Electric current occurs because of moving charges. Current may flow in a metal, a liquid, a gas, and a vacuum, but the charge carriers may vary from one medium to another. Commonly, charges are carried as a result of the movement of electrons or ions.
2. The structure of an atom can be represented as a positively charged nucleus with electrons moving in various orbits around this nucleus. The electrons in a particular orbit possess a certain level of energy. The smaller the orbit (closer to the nucleus), the lower the energy level.
3. Electrons can jump (make "transitions") from one orbit to another, or from one energy level to another. An electron may acquire additional energy from a source outside the atom (from light, heat, collisions, etc.) and jump to a higher energy orbit. On the other hand, an electron may lose energy (by radiation) and fall to a lower energy state -- make the transition to a smaller orbit.
4. The closer the electrons are to the nucleus, the more tightly they are bound to it; the difference in energy levels between small orbits being greater than the difference between larger orbits. The outermost electrons (called the "valence" electrons) are loosely bound to the nucleus. It would take only a relatively small amount of external energy to free them completely from the nucleus. The energy required to free a valence electron will be different for one type of atom than for another. For the metal germanium, only 0.7ev (electron-volts) is required.
5. The ability of a solid material to conduct electricity depends upon the availability of charge carriers. In some materials, the valence electrons are tightly bound to the nucleus and are not available as charge carriers. These materials conduct electricity poorly and are therefore used as insulators.
6. In other materials, like metals, the valence electrons are so loosely bound that they are practically free of the nucleus. Even a small applied electric field can get these electrons to move through the material. These materials are good conductors.

7. Still other materials fall between these two extremes; they are neither good insulators nor good conductors. Examples of such materials are germanium and silicon. They are semiconductors.

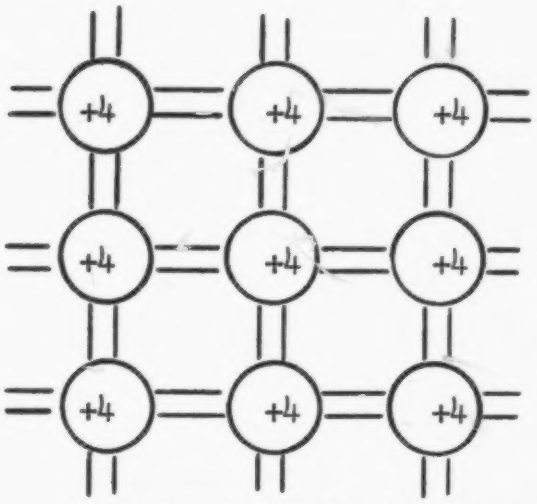
8. The number of free electrons in conductors is enormously greater than in semiconductors. At room temperatures:

   For conductors, approximately 1 electron per atom.

   For germanium, approximately 1 atom in 10^{10} is ionized.

   For silicon, approximately 1 atom in 10^{13} is ionized.

9. Pure Ge (Germanium) or Si (Silicon) atoms are arranged in a crystal lattice. Both atoms have 4 valence electrons in the outermost shell. This largest orbit would be complete if it contained eight (8) electrons; there are, therefore, 4 empty place in this outer shell. In a regular (uniform) crystal lattice, each atom "shares" an electron with each of its four neighboring atoms. The atoms that share each other's valence electrons thus form "covalent bonds." In this way each atom comes closer to having its outer shell filled.



A 2-Dimensional Model of a Crystal Lattice

Figure 9.1 is a two-dimensional presentation of a germanium lattice showing covalent bonds between adjacent atoms. These bonds hold the atoms fixed in space relative to each other and thus the collection of atoms forms a regular crystal lattice.

10. The atoms in a lattice, although fixed in position relative to each other, vibrate around their equilibrium positions. The degree of agitation is a function of temperature. As temperature increases the thermal agitation may be enough to break a covalent bond and an atom can become ionized by freeing an electron. (It takes .7ev of energy to free a valence electron in Ge.) Free electrons formed in this manner leave behind a hole, an empty position for another electron to fill. These holes effectively have a positive charge.

11. As free electrons move about in the crystal, they may encounter a hole and by "captured" by it. The electron then is recombined with an atom. These two tendencies -- ionization and recombination -- are opposite in nature. Under equilibrium conditions at a given temperature the rates of ionization and recombination are the same. In the process of ionization and recombination, as electrons move about, the holes appear to move as well. For all intents and purposes the holes can be considered as mobile charge carriers like electrons except that they have a positive sign.

12. In addition to pure crystals of a semiconductor, it is possible to have semiconductor crystals in which some foreign, impurity atoms are introduced deliberately. It is possible to increase the density of free electrons in a crystal by adding impurity atoms having a valence of +5 (five valence electrons in the outer shell). When this +5 material is incorporated in a germanium crystal, covalent bonds are again formed by sharing electrons with neighboring atoms, but now the fifth valence electron cannot be shared in any covalent bond. Since the outer layers of all atoms are effectively filled, the extra electron is not bound to any nucleus and is thus effectively free. In this way, there are free electrons without corresponding holes. The valence-5 atom "donates" a free electron and so is called a donor atom. (Actually it requires about .01ev of energy to free this electron compared with .7ev for pure germanium.) Since the charge carriers having a negative charge (electrons) are in greater number than those carrying a positive charge (holes), a semiconductor material having these characteristics is called N-type (Negative-type). The material is electrically neutral, however; there is no net charge. The process of adding the impurity atoms is called doping. In typical doped semiconductors, the fraction of impurity atoms is of the order of one part in 10^6.

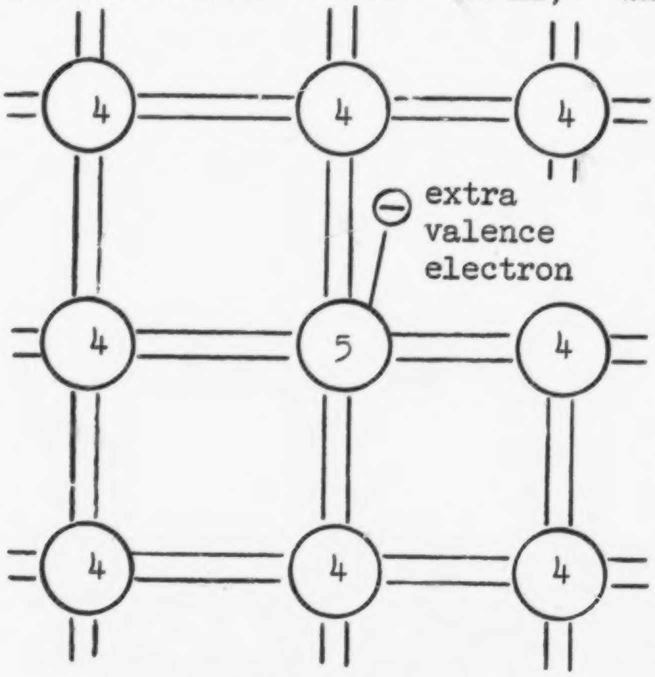


Fig. 12.1
N-Type Material

13. Another alternative is to include some impurity atoms with valence +3 (like gallium or indium). Each such atom again shares electrons with its neighbors to form covalent bonds. But now there aren't enough electrons to go around, so that a hole is formed. Any free electrons moving by can "fall into" this hole. Since the valence-3 atom "accepts" such electrons, it is called an acceptor atom. Each such impurity atom introduces an extra hole without a corresponding electron. Since the majority charge carriers in this type of material carry a positive charge, the material is called P-type.

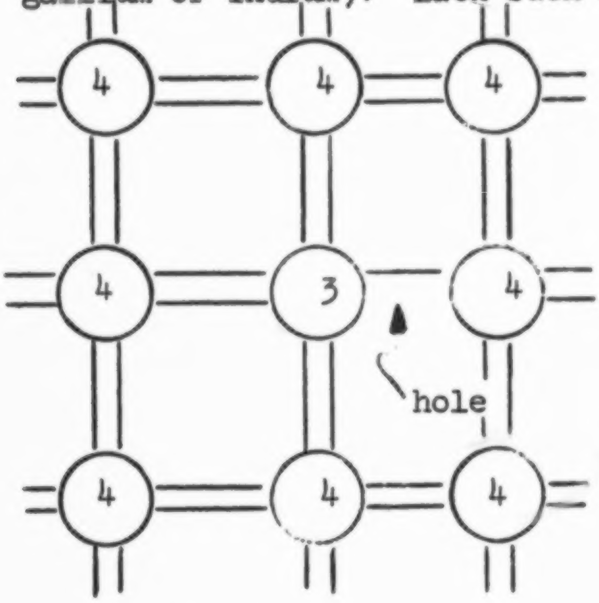


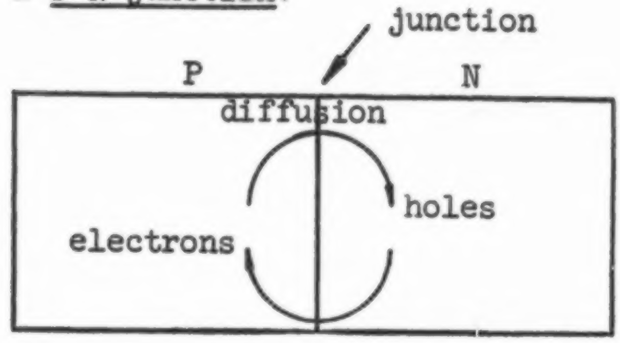
Fig. 13.1
P-Type Material

14. Charge carriers move through a semiconductor by two mechanisms: by diffusion and by drift. (a) Diffusion is the process connected with random motion due to thermal agitation. (b) Drift motion is due to an externally applied electric field and is superimposed on top of the random motion. At room temperature the diffusion velocity in crystals is of the order of 10^5 meters/sec., which is quite high. On the other hand, drift velocities are much smaller. For example, for N-type germanium the drift velocity is of the order of 1 meter/sec.

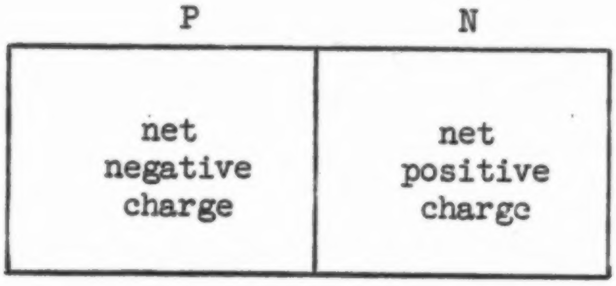
Summary. In doped semiconductor crystals, there are two sources of charge carriers: those due to thermal ionization of any atom in the crystal, and those resulting from the impurity atoms. In typical semiconductor devices, the density of impurity atoms is in the range 10^{15} to 10^{17} per cm^3. The density of charge carriers due to thermal ionization is only about 10^{13} per cm^3 for germanium and 10^{10} per cm^3 for silicon. Since each impurity atom leads to 1 charge carrier, it is seen that the number of charge carriers due to impurity atoms greatly exceeds that due to thermal ionization. Thus, for P-type material the majority carriers are holes and for N-type materials the majority carriers are electrons.

The P-N Junction Diode

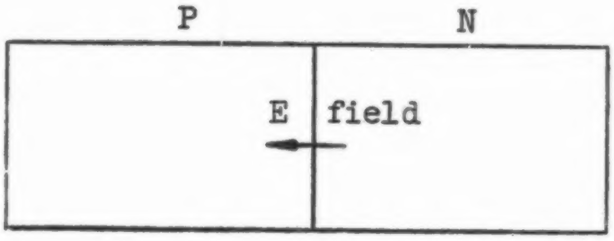
15. A piece of P-type material in contact with a piece of N-type material forms a P-N junction.



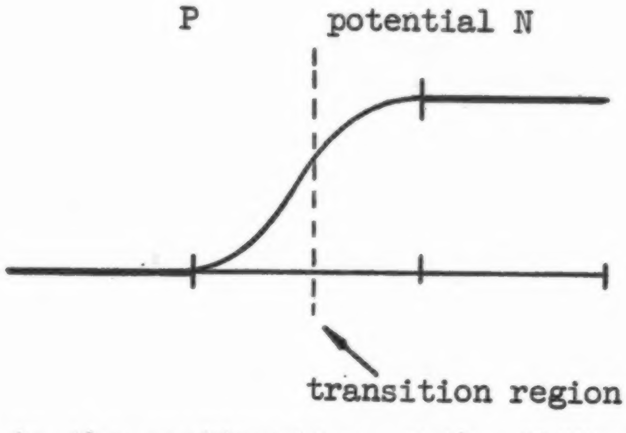
The charge carriers in each type diffuse into the other material. Thus, electrons (majority carriers in N type) diffuse to the left and holes (majority carriers in P-type) diffuse to the right



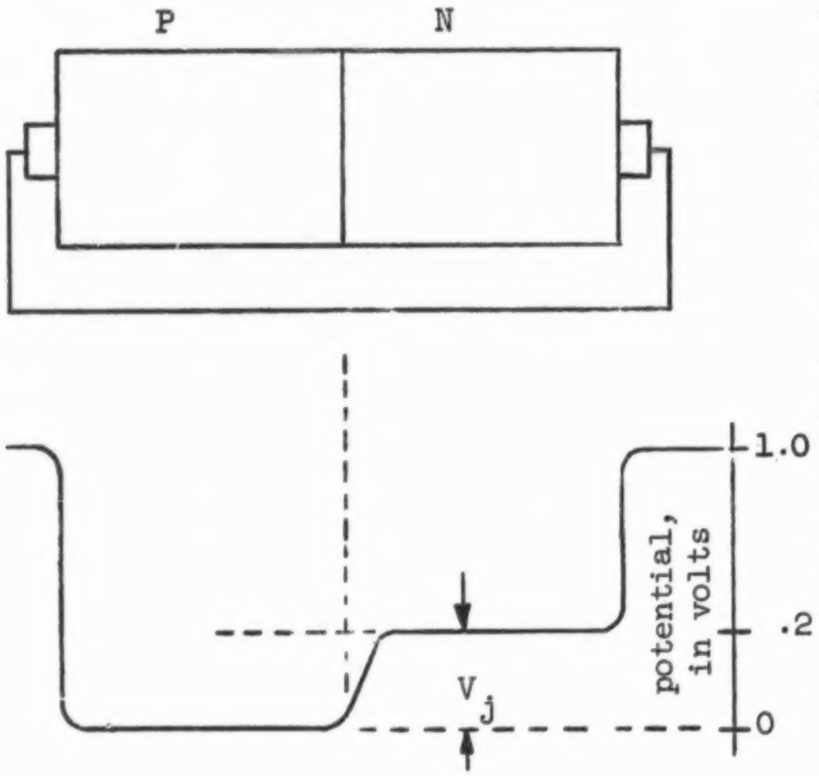
Electrons leaving the N region leave it positively charged and at the same time carry an excess negative charge into the P region. Holes leaving the P region add to this effect by leaving the P negatively charged and carrying a positive charge to the N region.



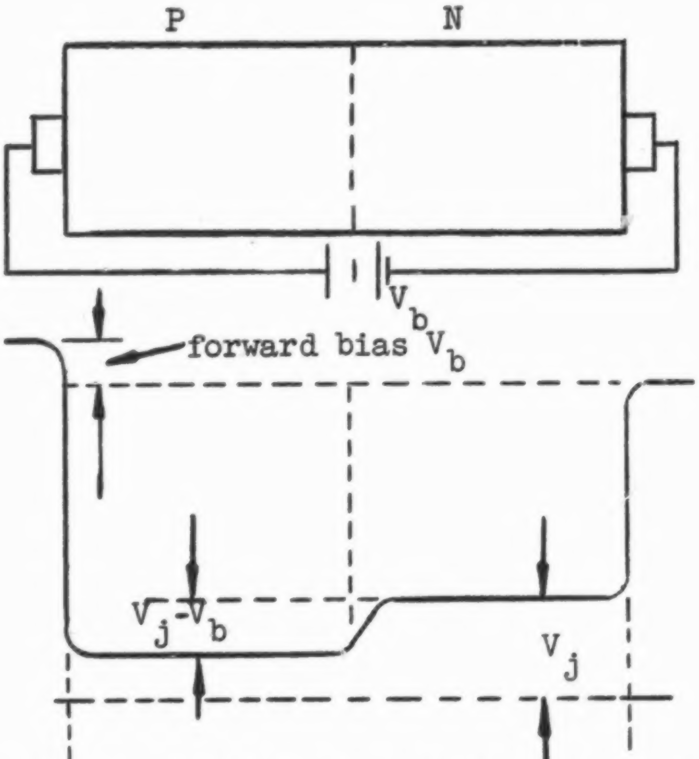
This separation of charges will produce an electric field which tends to inhibit further diffusion of charges.



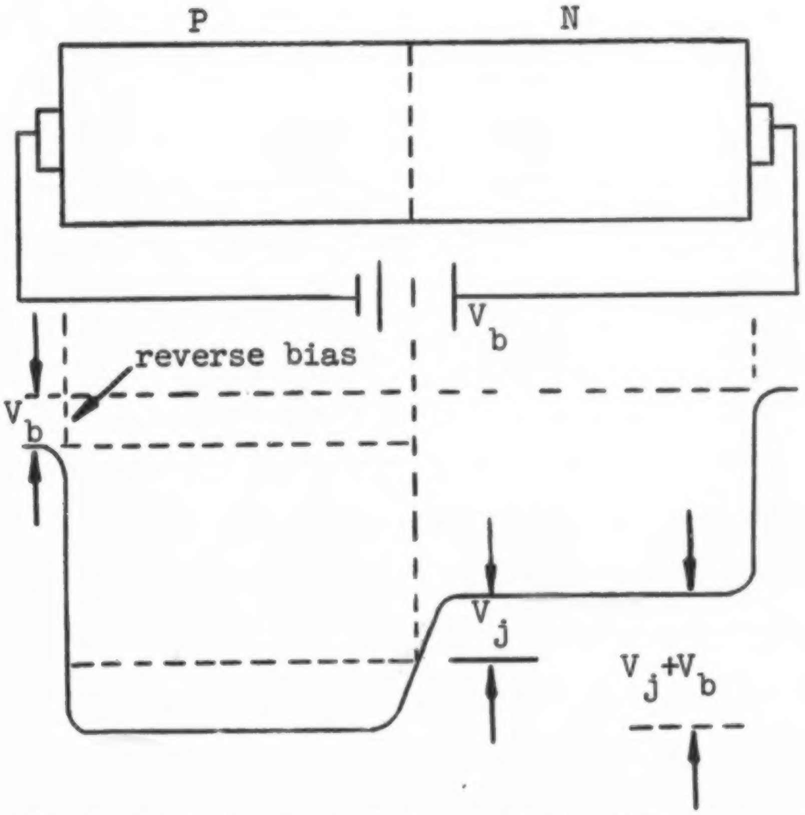
Because of the electric field at the junction, there will be a difference of potential across the junction between the two regions, as illustrated. The potential does not change abruptly when going from one region to the other, but changes gradually throughout a transition region surrounding the junction. (The width of this transition region is typically 1 micron = 10^{-6} meter.) The potential "hill" constitutes a barrier to the continuation of the diffusion process. The height of the "hill" is determined by an equilibrium between thermal ionization on both sides of the junction and diffusion of carriers across the junction. In equilibrium, there is no net current flow across the junction.



The device formed in this manner is a junction diode. In order to connect the diode in a circuit, metal leads are attached to the two ends. Suppose the ohmic voltages at the lead junctions are negligible. Now let the two leads be short circuited, so that the potentials of the two ends are the same. The diagram shows the resulting variation of potential throughout the device. The values shown are typical. There will be no current flowing in the short circuit.



Now let a battery be connected across the diode with the positive terminal connected to the P-region. The potential of the P-region is increased by the battery voltage V_b while the potential of the N-region remains the same. Thus, the potential hill at the junction is reduced by the voltage V_b, becoming $V_j - V_b$. The equilibrium which kept the net current across the junction is thus disturbed. Since the potential barrier has been lowered, it is easier for the majority carriers (electrons in the N-region and holes in the P-region) to diffuse in their natural direction. This flow will not now cause an increase in the junction potential since the charge carriers do not accumulate, they continue flowing around the closed circuit. The variation of potential now takes the form shown. The diode is said to be forward-biased.

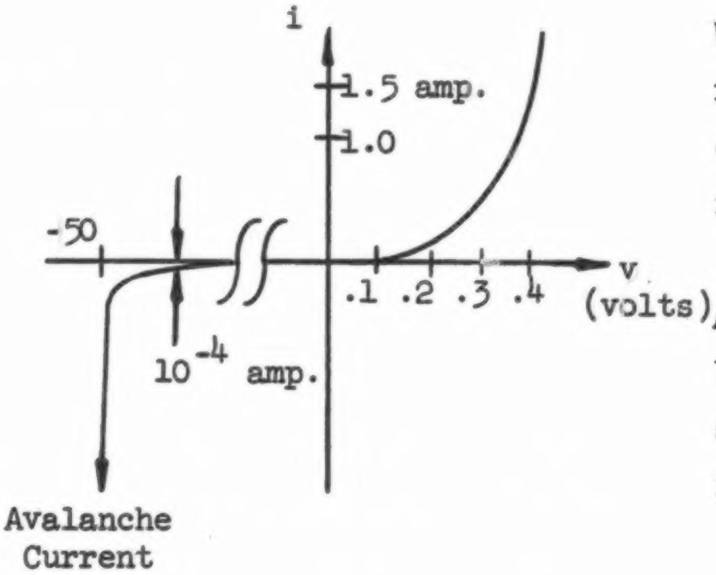


If the polarity of the biasing voltage is reversed, the diode is said to be reverse-biased. Now the potential of the P-region is reduced by V_b while that of the N-region is unchanged. Hence, the height of the potential hill is increased to $V_j + V_b$. The majority carriers will now be hindered in moving in their natural direction. However, there is always some thermal ionization leading to minority carriers in each region (electrons in P-region, holes in N-region). For these minority carriers, the direction of the electric field is just right for helping their diffusion. Thus, there will be a certain amount of reverse current. The magnitude of the reverse current compared with that of the forward current is approximately as the density of minority carriers to that of majority carriers. This is of the order of 10^{-2} to 10^{-4} for germanium diodes and 10^{-5} to 10^{-7} for silicon diodes. Since the origin of reverse current is thermal ionization, this current is sensitive to temperature.

Zener and Avalanche Effects

16. If the reverse bias applied to a diode is increased (V_b is made larger), the potential barrier will increase in height and so will the magnitude of the electric field. As the field increases it will become so high that electrons can be torn away from the covalent bonds, thus creating hole-electron pairs in the transition region. Although these pairs are not created by thermal ionization, their presence will lead to an increase in reverse current. In fact, this increase will be a large one. This effect is called the Zener effect.

There is another possibility for increasing the reverse current without the breaking of covalent bonds by the Zener effect. The carriers constituting the small reverse current will have collisions with atoms in the crystal lattice. If the electric field in the transition region becomes high enough, these carriers may gain sufficient energy between collisions to knock electrons out of the covalent bond, again creating an electron-hole pair which contributes to the reverse current. This effect is cumulative since the carriers so created can also lead to ionizing collisions. The result is called the avalanche effect.



Whichever mechanism is involved, the result is a great increase in reverse current which takes place without much further increase in reverse bias voltage.

A sketch of typical diode current against bias voltage is shown. (Note that the scales on the positive axes are different from those on the negative axis.)

A REVIEW TEST

Test yourself on your retention of the material you have just read. If you are unsure of any answers, go back and re-read the related paragraphs.

1. In a P-type semiconductor material consisting of germanium doped with an impurity material, the valence of the impurity atoms is:
    - a) 4
    - b) 3
    - c) 5
    - d) either 3 or 5

2. In an N-type semiconductor material consisting of germanium doped with an impurity material, the valence of the impurity atoms is:
    - a) 4
    - b) 3
    - c) 5
    - d) either 3 or 5

3. In a P-type semiconductor the majority carriers are holes.
    - a) True
    - b) False

4. In an N-type semiconductor the minority carriers are holes.
    - a) True
    - b) False

5. A P-type semiconductor has an excess of holes over free electrons. The material as a whole is:
    - a) negatively charged
    - b) neutral
    - c) positively charged

6. An N-type semiconductor has an excess of free electrons over holes. The material as a whole is:
   - a) negatively charged.
   - b) neutral.
   - c) positively charged.

7. Charge carriers move through semiconductors as a result of:
   - a) drift due to electric fields.
   - b) diffusion due to random processes.
   - c) both a) and b)
   - d) none of the above

8. Forward bias on a P-N diode:
   - a) increases the potential hill at the junction.
   - b) decreases the potential hill at the junction.

9. Reverse bias on a P-N diode:
   - a) increases the potential hill.
   - b) decreases the potential hill.

10. A P-N diode will not break down, no matter how high the reverse bias.
    - a) True
    - b) False

11. The avalanche effect is the result of:
    - a) operating the device at improper temperatures.
    - b) applying excessive reverse bias.
    - c) doping the germanium incorrectly.

Chapter 9

TRANSISTOR AMPLIFIERS

The Transistor

A transistor is a semiconductor device in which two junction diodes are effectively placed back to back, the net result consisting of three layers of material, as shown in Fig. 1(a). The center of the "sandwich" is P-type material while the outer layers are N-type. The resulting device is called an N-P-N junction transistor. The center slice is called the base; the lower N-region is called the emitter and the upper N-region, the collector.

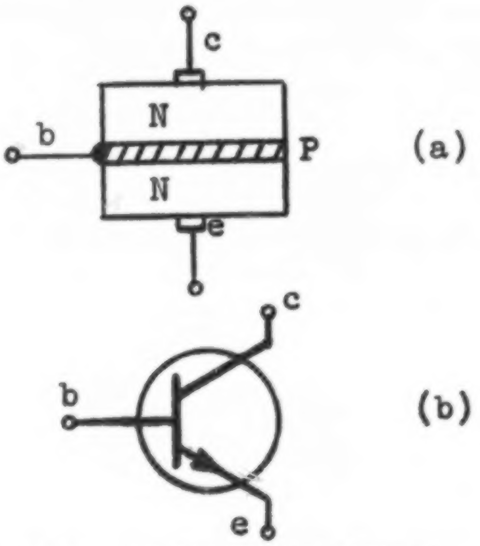


Fig. 9-1 N-P-N Transistor

Rather than drawing the pictorial representation of a transistor shown in Fig. 1(a), a standard symbol is used as shown in Fig. 1(b). The emitter is distinguished from the collector by means of the arrow. The direction of the arrow is the direction of forward current.

Remember from the discussion of P-N junctions that the junction is forward-biased when the potential of the P-region is made positive relative to that of the N-region; that is, from the base to the emitter. The arrow on the transistor symbol will, thus, have the direction shown.

Forward current across a junction generally has contributions from two sources: holes (which are the majority carriers in the P-region) moving from P to N and electrons (which are the majority carriers in the N-region) moving from N to P. In the junction transistor the forward current actually consists largely of electrons injected from the emitter to the base. (Remember that negative charge flow is opposite to the direction of the current.) This is achieved by making the base region physically very small and also making the concentration of impurity atoms in the base small compared with that of the emitter. The forward current flow is facilitated by biasing the base-emitter junction (which will be more simply referred to as the emitter junction) in a forward direction; that is, making the base positive relative to the emitter

In the base region (P-type material), the electrons are minority carriers; as they diffuse through the base, some of them recombine with holes, which abound there. If the base layer is thin enough most of the electrons will reach the base-collector junction without recombining. If we now arrange to create a field drawing these electrons across the base-collector junction, we will, in effect, cause an electron flow from emitter, through the base, to the collector (that is, cause a current to flow from collector to emitter). Such a field is obtained by biasing the base-to-collector junction in the reverse direction; that is, making the base negative relative to the collector. The appropriate biases are shown in Fig. 2

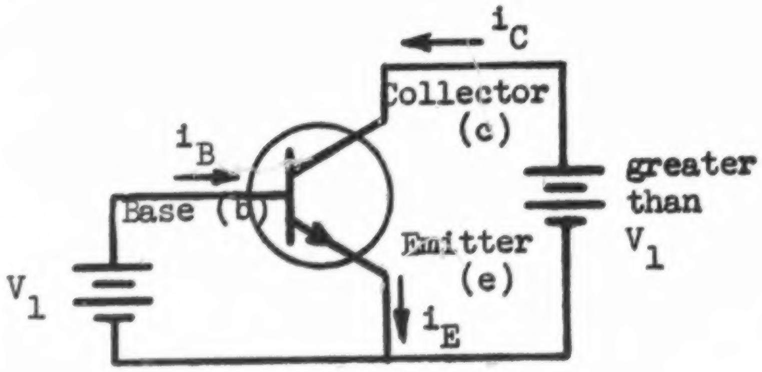
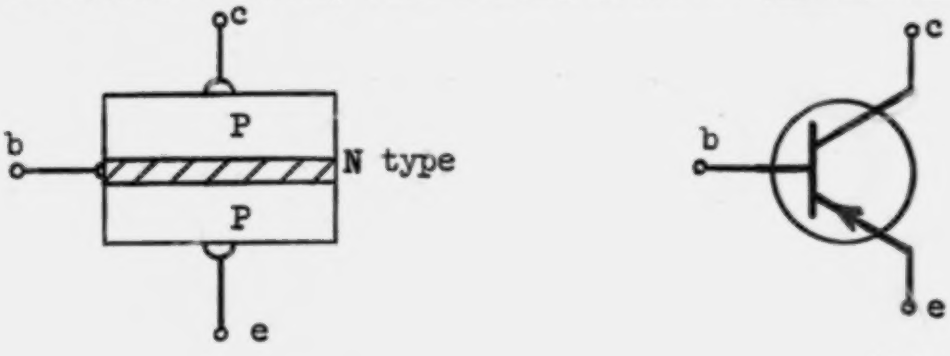


Fig. 9-2

With such an arrangement, the collector current i_c is approximately equal to the emitter current i_e. There is some base current which supplies the few holes that recombine with electrons in the base, and also supplies the holes which diffuse across the emitter junction. By proper choice of the densities of impurities in the base and emitter regions, and by making the base very thin, the hole current across the emitter junction can be made much smaller than the electron current across this junction. Hence, a small base current will control a large collector current, thus making possible current amplification.



(a) (b)

Fig. 9-3. P-N-P Transistor

This is a basic purpose of the transistor.

It is also possible to make a junction transistor of the P-N-P variety by reversing the materials of the N-P-N transistor. A base of N-type semiconductor is sandwiched between two pieces of P-type material, as shown in Fig. 3(a). Forward current now consists of holes moving across the emitter junction from emitter to base. The arrow on the emitter in the symbol for the P-N-P transistor shown in Fig. 3(b) indicates this. Note that the arrow is always directed from P to N, so the direction of the arrow alone is enough to distinguish between N-P-N and P-N-P transistors.

In order to bias the emitter junction of a P-N-P transistor in the forward direction, it is necessary to make the base negative relative to the emitter. Similarly, to give the collector junction a reverse bias it is necessary to make the base positive to the collector. These are just the opposites of what was needed in an N-P-N transistor. In a circuit, a P-N-P transistor can replace a similar N-P-N transistor provided the polarities of all bias batteries are reversed.

The Transistor Amplifier

Figure 4 is a diagram of a basic transistor amplifier. The arrow on the emitter shows the transitor to be a N-P-N transistor. The batteries V_{CC} and V_{BB} provides biasing and the source v_s provides the signal which is to be amplified.

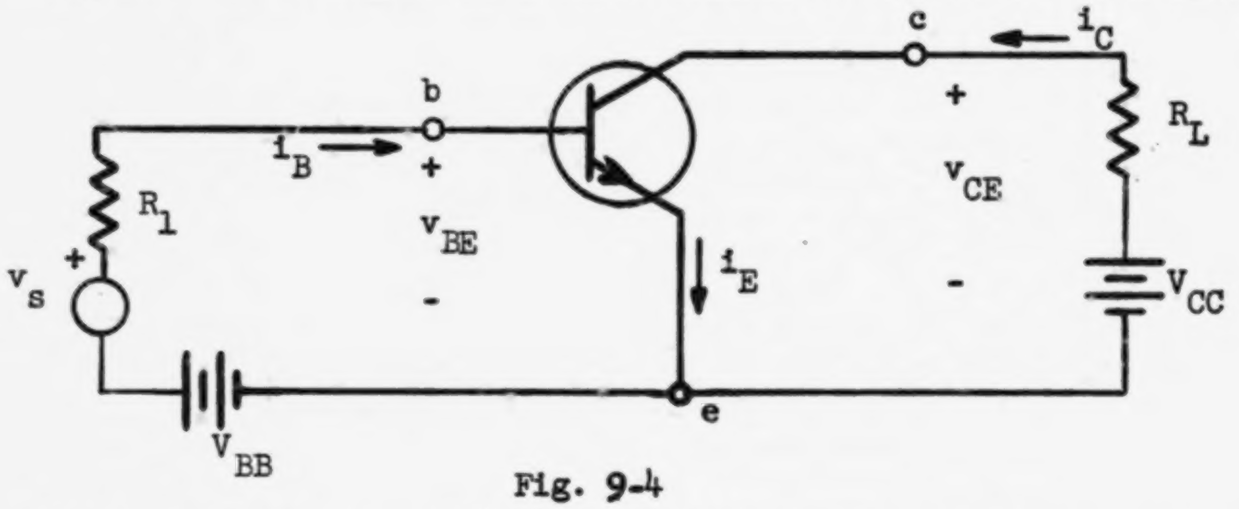


Fig. 9-4

The behavior of the transistor can be described in terms of its terminal voltages and currents. By Kirchhoff's voltage law the three possible voltages across the transistor terminals are related by

$$v_{BC} = v_{BE} - v_{CE} \tag{9-1}$$

Thus, giving two of the voltages will determine the third; only two of them are independent. Similarly the three currents are related by

$$i_E = i_B + i_C \tag{9-2}$$

and again only two of the currents are independent.

Now, in order to determine how the two voltages and the two currents are related, we vary some of them and measure the variations in the others.

(It is not important how they are varied; it can be done by changing V_{CC} or V_{BB}.) Typical curves obtained are shown in Fig. 5. To get a complete description of the current-voltage relationships two families of curves are required.

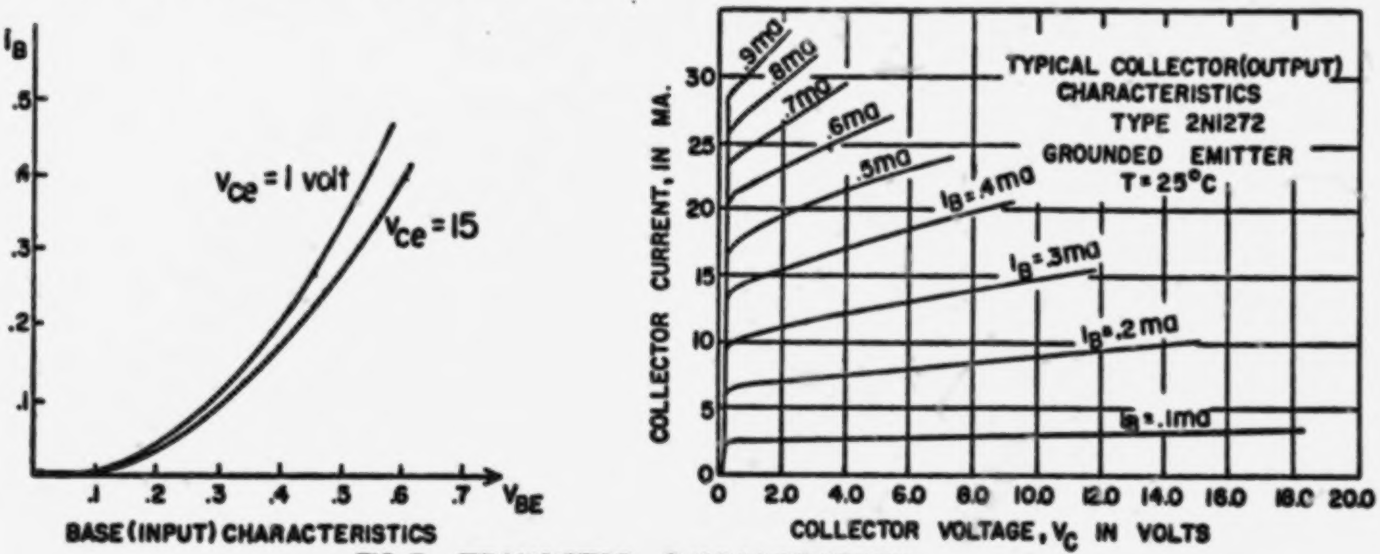


FIG. 5 TRANSISTOR CHARACTERISTICS

In one family of curves the output (collector to emitter) voltage is held constant at some value and a plot of the variation of base current with input (base to emitter) voltage is made. This is repeated with v_{CE} held at other constant values. Figure 5(a) shows that there isn't much variation in this input characteristic as v_{CE} varies from 1 to 15 volts.

The second family of curves is obtained by holding the base current constant and plotting variations of output (collector) current with output (collector to emitter) voltage. The result is shown in Fig. 5(b). Note that even with zero base current there is some output current which is approximately constant for all values of v_{CE}. Note the common features of this family of collector characteristic curves. Except for low values of v_{CE}, the curves are almost horizontal lines with small upward slopes. They are almost evenly spaced for equal increments in base current.

In order to analyze the operation of the amplifier shown in Fig. 4, we write two loop equations around the base-emitter loop and the collector-emitter loop.

Thus,

$$v_{BE} - V_{BB} - v_s + R_1 i_B = 0 \tag{9-3}$$

$$V_{CC} - v_{CE} - R_L i_C = 0 \tag{9-4}$$

Solving these for i_B and i_C leads to

$$i_B = \frac{V_{BB} - v_{BE}}{R_1} + \frac{v_s}{R_1} \tag{9-5}$$

$$i_C = \frac{V_{CC}}{R_L} - \frac{v_{CE}}{R_L} \tag{9-6}$$

These are two equations in four unknowns. Two more relationships among the unknowns are required for a solution. And we do have two such relationships given by the transistor characteristics in Fig. 5. The input characteristic shows that v_{BE} is small; it is normally small (one tenth or less) compared with V_{bb}. Hence, to a first approximation, it can be neglected in Eq.(9-5) and this equation can be rewritten as

$$i_B = \frac{V_{BB}}{R_1} + \frac{v_s}{R_1} = I_B + i_s \tag{9-7}$$

where $i_s = v_s/R_1$ is the signal current and, it is assumed that v_s is a time-varying voltage whose average value is zero; $I_B = V_{BB}/R_1$ is, hence, the average value of the base current.

Equation (9-6) is the equation of a straight line whose variables are the coordinates of the collector characteristic shown in Fig. 5(b). This straight line can be superimposed on the collector curves, as shown in Fig. 6.

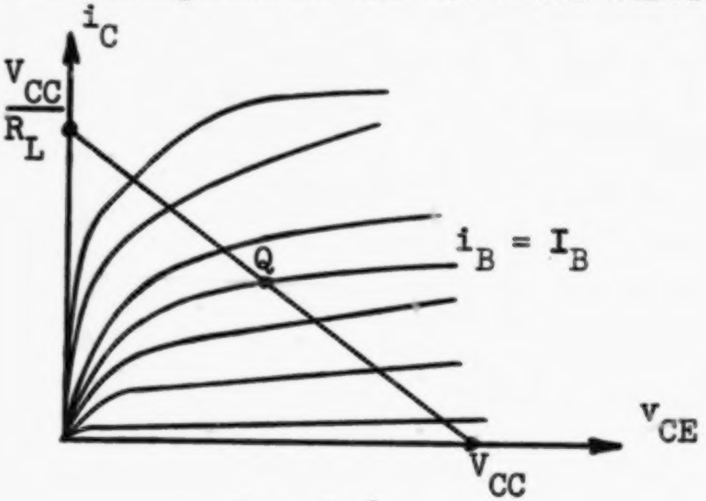
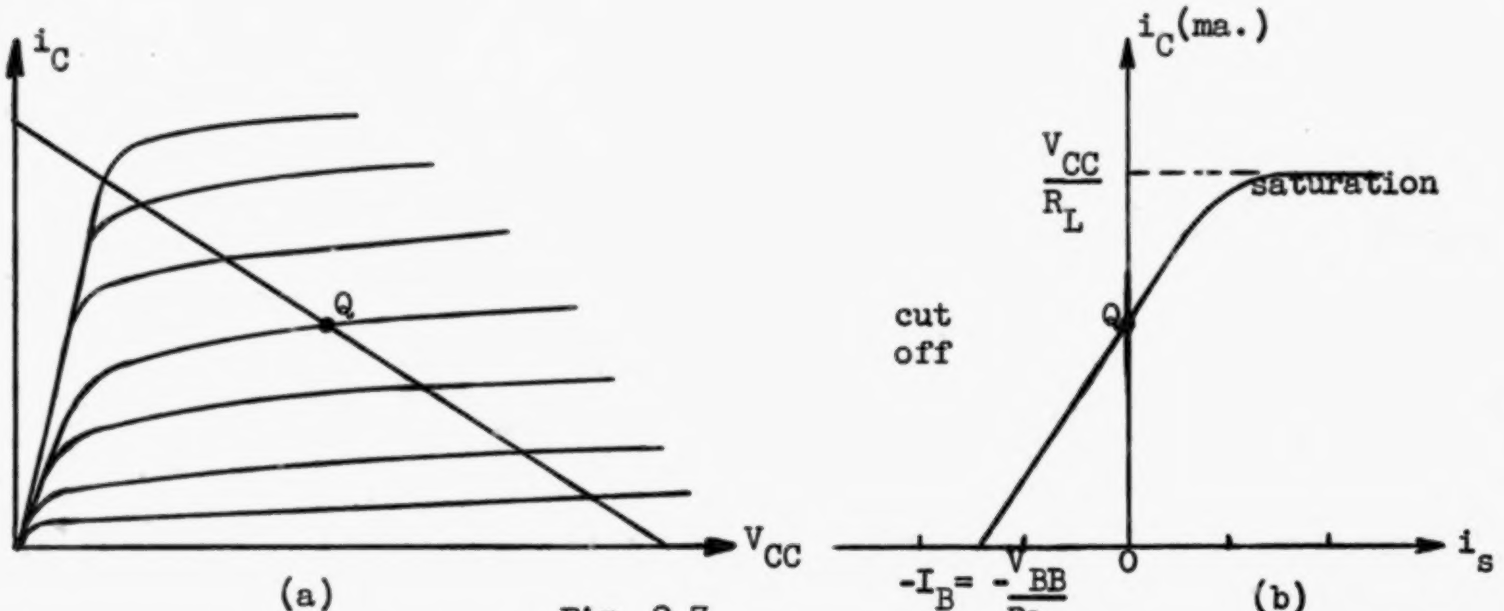


Fig. 9-6

This line is called the load line. (You will recall a similar graphical analysis of a diode circuit.) The signal voltage v_s is time varying and will cause the base current i_B to vary. For any instantaneous value of i_B, the corresponding value of i_C and V_{CC} can be determined by noting the coordinates of the intersection of the load line and the appropriate i_B curve. Thus, as the signal voltage varies, the collector current and v_{CE} will both change, but in such a way that their values will always lie on the load line.

When the signal current ($i_s = V_s/R_1$) is zero, the base current i_B will have its average value I_B (Eq. 9-7). The intersection of the load line with the i_B-curve corresponding to this average value of base current I_B is called the operating point and is labeled Q in Fig. 6.

It is of interest to determine the variations in collector (output) current as a function of the signal current i_s. This can be done by assuming positive and negative values of i_s, adding I_B (the average base current) to get i_B, then reading the corresponding value of i_C from the intersection of the load line with the corresponding curve. The result is graph shown in Fig. 7.



(a)

(b)

Fig. 9-7

Over a substantial part of the range this curve is approximately linear indicating that the output (collector) current is approximately proportional to the input (signal) current, the proportionality constant being the slope of the curve.

When the signal current becomes sufficiently negative, the collector is cut off and no collector current flows. At the other end, for large positive values of signal current, the collector current reaches a maximum and from then on it is saturated. If an input signal is to be reproduced faithfully, its amplitude is limited by these considerations. In the design of the amplifier the operating point must be so chosen that the point $i_s = 0$ falls near the center of the linear portion of the curve of Fig. 7, assuming that the signal current has equal positive and negative peaks, like a sinusoid.

Example: Let the curves in Fig. 8 be the approximate collector characteristics of an N-P-N junction transistor to be used in the amplifier circuit of Fig. 9. The bias battery voltages are $V_{CC} = V_{BB} = 15$ volts and $R_1 =$ 75 K ohms. It is required to find the value of R_L which will cause v_{CE} at the operating point to be 10 volts. It is also required to determine the peak value of signal voltage which can be amplified without distortion.

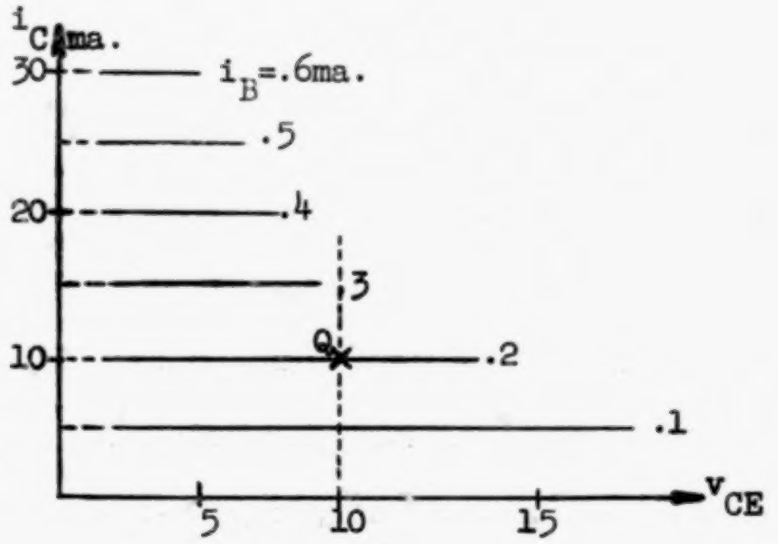


Fig. 9-8

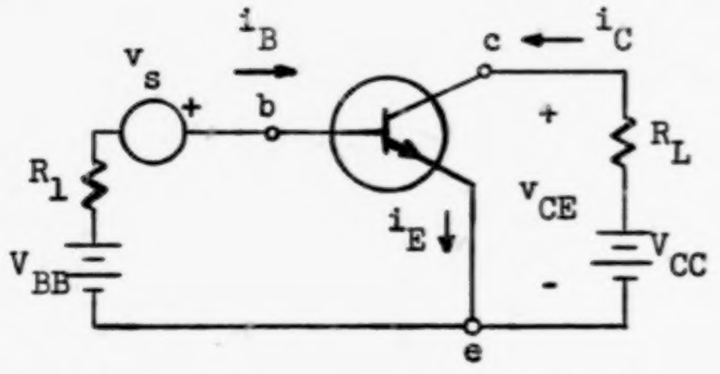


Fig. 9-9

Making the approximation that v_{BE} can be neglected relative to V_{BB} (15 volts), the average value of base current will be $I_B = 15/75000 = 0.2$ ma. The coordinates of the operating point will then be $v_{CE} = 10$, $i_C = 10$ as shown by the X in Fig. 8. Since another point on the load line is $v_{CE} = 15$, $i_C = 0$, the load line can now be drawn, giving an intercept on the i_C axis of 30 ma. The result is shown in Fig. 10. The slope of the load line is found to be - 30 x 15 3/15 = - 1/500. From Eq. (9-6) the slope is $-1/R_L$; hence, R_L = 500 ohms.

Assume that the input signal v_S has equal positive and negative peaks (for example, the output of a phonograph pickup would be such a symmetrical wave). This signal would cause i_B to vary along the load line, around the operating point Q. If the signal were large enough, the base current could go negative, driving i_C to cut-off (Fig. 7b). Under this condition, the output of the amplifier would be distorted; a portion of the input signal v_S would be lost.

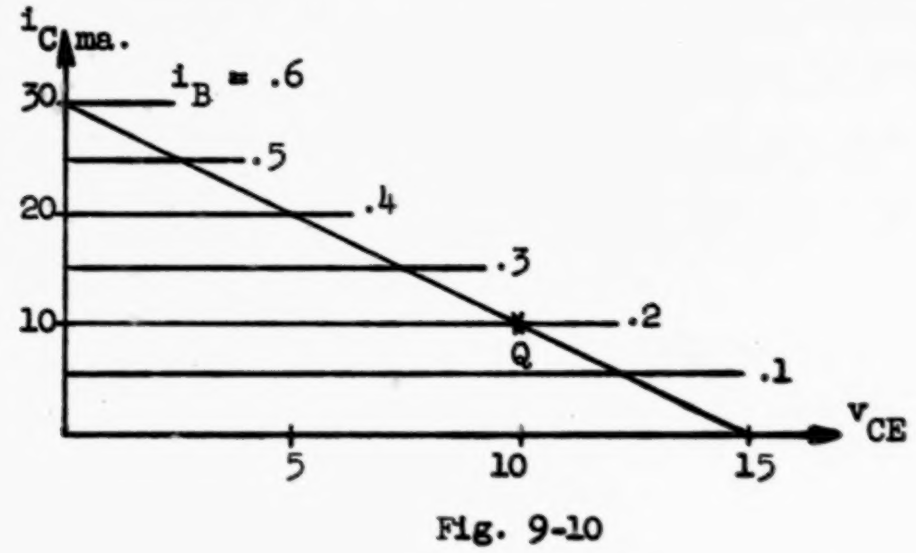


Fig. 9-10

Since from Eq. (9-7) the total base current equals the average value plus the signal current, the *negative* peak of the signal current should not exceed 0.2 ma, since $I_B = 0.2$. This means that the largest base current will be .4 ma.

The characteristics show that the output current is linearly related to the base current over a base current range from 0 to .6ma. If the operating point can be moved to the *center* of this range, then a signal having a larger peak-to-peak value could be amplified without distortion.

Find the value of R_1 and R_L required to move the operating point to $i_B = .3$ma, $N_{CE} = 7.5$ volts

Ans: $R_1 = 50$K
$R_L = 500$ ohms

Symbols for Circuit Variables

The symbols we have been using for the voltages and currents in a transistor amplifier(such as in Fig. 11, below) have been lower case (small) v and i with upper case (capital) letter subscripts These capital subscripts refer to the total instantaneous values. In discussing the operation of amplifiers we will have occasion to talk about both the average values of voltages and currents as well as their instantaneous values. Furthermore, these voltages and currents can be written as the sum of their average value plus an alternating component whose average value is zero. So there is need for other symbols to designate these alternating components.

The following conventions will be used.

1. Capital letters with repeated capital subscripts will represent supply or bias sources. Thus, V_{CC} is the collector supply voltage and V_{BB} is the base bias voltage. (Note the two subscripts are the same.)

2. Capital letters with capital subscripts, either single or two different letters, will represent average values. Thus, I_B is the average base current and V_{VE} is the average collector to emitter voltage.

3. Lower case letters with capital subscripts, wither single or double, will represent total instantaneous values. Thus, v_{BE} is the total base to emitter voltage; an average value plus an a-c component. (Note, in case of double subscripts, the two subscripts are different.)

4. Lower case letters with lower case subscripts will represent instantaneous values of alternating components. Thus, i_e is the alternating component of emitter current and v_{be} is the alternating component of base-to-emitter voltage (their average values will be zero).

5. Capital letters with lower case subscripts will refer to the rms values of alternating components. Thus, I_e will be the rms value of i_e, and V_{be} will be the rms value of v_{be}.

Self-Test. Can you select the correct answers to the following questions? There may be more than one correct answer. (Check your answers against those given in the footnote, below.)

1. The term "v_{CE}" designates
   - A. The instantaneous voltage from collector to emitter.
   - B. The sum of V_{CC} and i_eR_L.
   - C. The sum of $V_{CE} + v_{ce}$.
2. The term V_{be} refers to
   - A. The average value of base-to-emitter voltage.
   - B. The rms value of base-to-emitter voltage.
   - C. The value of I_eR_1.
3. The average value of collector-to-emitter voltage is
   - A. V_{CE}
   - B. $v_{CE} - v_{ce}$
   - C. V_{ce}

Answers: 1. A,C; 2. B,C; 3. A,B.

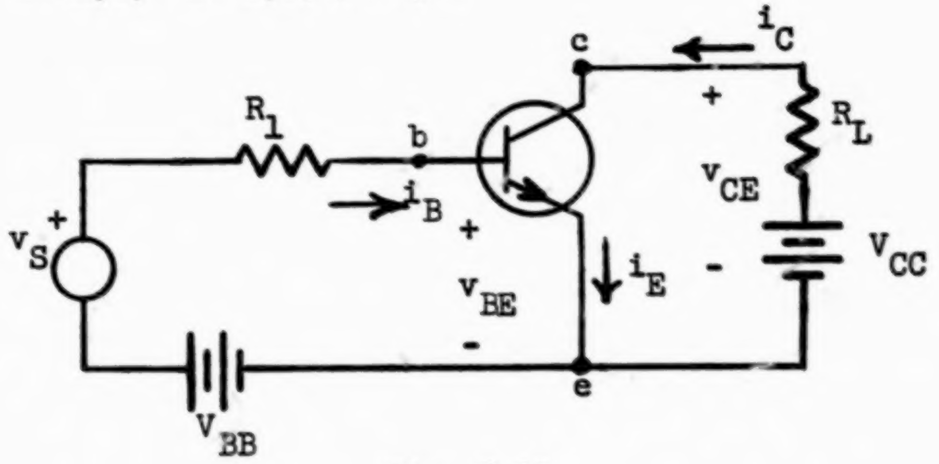


Fig. 9-11

Power in the Collector Circuit

Having established the convention for the notations to be used, let us concentrate on the collector circuit of Fig. 11 -----. the loop formed by the collector and emitter terminals of the transistor, the supply voltage V_{CC} and the load resistance R_L. Interest is to be focussed on three aspects of power, (1) the power supplied by the source, (2) the power dissipated in the transistor, and (3) the power dissipated in the load. We will more specifically, be interested in the average power supplied by or dissipated in each component in the collector circuit.

The first step will be to write expressions for the pertinent voltages and currents in terms of average and alternating components. Thus

$$i_C = I_C + i_c \tag{9-8}$$

$$v_{CE} = V_{CE} + v_{ce} \tag{9-9}$$

We assume that the alternating components have zero average values.

The instantaneous values of the power supplied by the source (p_{CC}), the power dissipated in R_L(p_L), and the power dissipated in the collector (p_C) are

$$p_{CC} = V_{CC} i_C = V_{CC} I_C + V_{CC} i_c \tag{9-10}$$

$$p_L = R_L i_C^2 = R_L (I_C + i_c)^2 = R_L I_C^2 + R_L i_c^2 + 2R_L I_C i_c \tag{9-11}$$

$$\begin{aligned} p_C &= v_{CE} i_C = (V_{VE} + v_{ce})(I_C + i_c) \\ &= V_{CE} I_C + v_{ce} i_c + V_{CE} i_c + I_C v_{ce} \end{aligned} \tag{9-12}$$

Let us represent the average powers by capital P's and note that the average value of the alternating components are zero. We now take the average value of each of the last three expressions. We find:

$$P_{CC} = V_{CC}I_C \tag{9-13}$$

$$P_L = R_L I_C^2 + \text{average value of } R_L i_c^2$$

$$= R_L I_C^2 + R_L I_c^2 \tag{9-14}$$

$$P_C = V_{CE}I_C + \text{average value of } v_{ce}i_c$$

$$= V_{CE}I_C + V_{ce}I_c \tag{9-15}$$

Thus, the average power delivered by the supply source is not influenced by the presence or absence of a signal. On the other hand, both the average load power and the average power dissipated in the transistor are dependent on the signal. Note from Eq. (9-14) that the average load power has contributions from a dc component $R_L I_C^2$ and from an ac component, $R_L I_c^2$. (Remember that I_c is the rms value of the ac component of the collector current.) The same is true of the average collector power (dissipated in the transistor).

The three average powers can be related by noting that

$$v_{CE} = V_{CC} - R_L i_C \tag{9-16}$$

When this is used in Eq. (9-12) instead of the alternative expression for v_{CE}

given in Eq. (9-9), the result for the instantaneous collector power becomes

$$P_C = V_{CC} i_C - R_L i_C^2 = V_{CC}(I_C + i_c) - R_L(I_C + i_c)^2$$

Taking the average value leads to

$$P_C = \underbrace{V_{CC} I_C}_{\substack{\text{average power}\\ \text{supplied by}\\ \text{source}}} - \underbrace{(R_L I_C^2}_{\substack{\text{dc component}\\ \text{of average}\\ \text{load power}}} + \underbrace{R_L I_c^2)}_{\substack{\text{ac component}\\ \text{of average}\\ \text{load power}}} \qquad (9\text{-}17)$$

or

$$P_C = P_{CC} - P_L \qquad (9\text{-}18)$$

Under quiescent (no-signal) operating conditions, the average load power is just $R_L I_C^2$. With a signal, this average load power increases to $R_L(I_C^2 + I_c^2)$. Since the average power delivered by the source is constant, this means that, the average power dissipated in the transistor decreased with an increasing signal level. The maximum collector dissipation occurs with no signal. If we label this collector dissipation with no signal as P_{CO}, the collector dissipation can be expressed in terms of:

$$P_C = P_{CO} - R_L I_c^2 \qquad (9\text{-}19)$$

where

$$P_{CO} = I_C(V_{CC} - R_L I_C)$$

Example

Problem An N-P-N junction transistor is used in the amplifier circuit of Fig. 11. The transistor has a maximum permissible power dissipation of 2.3 watts.

The source voltages are $V_{CC} = 10$ and $V_{BB} = 14$ volts and the resistors have the values $R_1 = 2000$ and $R_L = 10$ ohms. The collector characteristics are approximately horizontal lines with $i_c = 60i_B$. A sinusoidal voltage signal v_s having an amplitude of 12.5 volts is applied. Find the average powers P_C, P_L and P_{CC} both under quiescent conditions and with the signal. Note whether or not the permissible dissipation is exceeded in either case.

Solution.

We first find the average base current to be $I_B = 14/2000 = 7$ma, assuming v_{BE} to be small compared with 14. The average collector current is then $I_C = 60 \times 7 = 420$ma.

The average power delivered by the source will be $10(.42) = 4.2$ watts. Under quiescent conditions the power dissipated in the load will be $10(.42)^2 = 1.76$ watts. Hence, the collector dissipation will be $4.2 - 1.76 = 2.44$ watts, which exceeds the maximum permissible.

The rms value of the current in the load resistor will be $\frac{12.5}{\sqrt{2}} \cdot \frac{1}{2000} \cdot 60 =$ .265 ma. Hence, the additional average load power will be $10(.265)^2 = .70$ watt giving a total average load power of 2.46. The collector dissipation is now reduced to $4.20 - 2.46 = 1.74$ watts, a safe value.

As the dynamic operating point (to be distinguished from the quiescent operating point, with no signal) moves up and down the load line, the instantaneous power dissipated in the collector changes. It is useful for the circuit designer to have a curve on which the collector dissipation is constant. Such a curve is obtained by setting p_C in Eq. (9-12) equal to a constant. In particular, if the constant is the maximum permissible dissipation, labeled P_{dm}, we will get

$$v_{CE} i_C = P_{dm} \qquad (9\text{-}20)$$

Since P_{dm} is a constant, this is the equation of a hyperbola. It is shown in Fig. 9-12 superimposed on a set of collector characteristics. For a given application, the amplifier should be so designed that the load line is tangent to or falls below this maximum dissipation hyperbola.

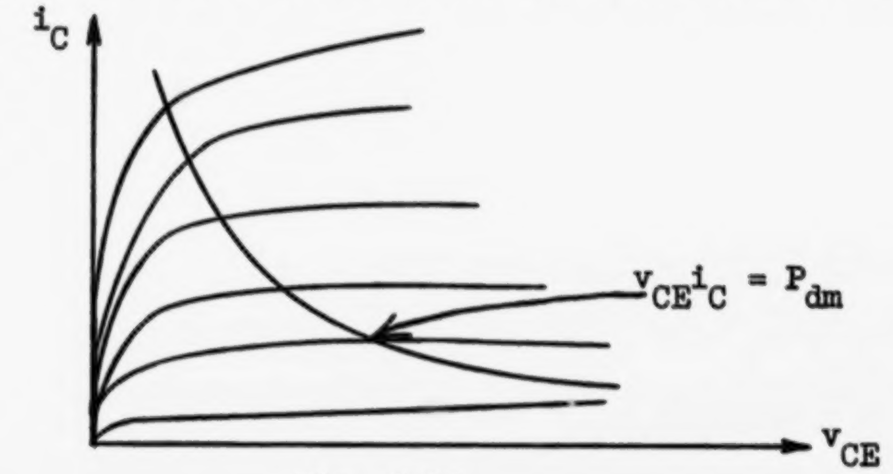


Fig. 9-12

Models of a Transistor

There are two ways in which one can proceed to obtain models, or equivalent circuits, of a transistor to use in assessing its behavior when it is connected in a network. One approach could be to study the transistor's physical principles and to arrive at equations expressed in terms of its terminal voltages and currents; these equations would contain expressions for such physical parameters as concentrations of holes and electrons in the P-region and N-region, diffusion lengths for holes and electrons, the charge of an electron, etc. These equations could then be related to an equivalent circuit. In this text, we shall not study the pertinent physical principles in such detail as will permit us to carry out such an analysis.

We have, however, discussed some of the elementary principles sufficiently to have a qualitative appreciation of some properties of a transistor. Consider again the diagrams of Fig. 9-1 and Fig. 9-2 and remember that a P-N junction is forward-biased when the P-region is made positive relative to N-region, and is reverse-biased when the opposite is true. With the biases shown in Fig. 9-2, the emitter junction is forward-biased and the collector junction reverse-biased. The electrons which are injected into the base from the emitter diffuse through the base region without much recombination, as already discussed. Because of the reverse-biased collector junction, there is a strong field pulling these electrons from the base into the collector. Thus, the collector current is almost equal to the emitter current.

Let us write

$$i_C = \alpha i_E \tag{9-21}$$

when α is an amplification factor whose value is less than 1, but very close to 1. (A typical value is $\alpha = .98$). Actually, for the amplifier connection we have been discussing, we are more interested in the amplification relating i_C to i_B. Such a relationship can be obtained by solving Eq. (9-21) for i_E and substituting into Eq. (9-2). The result will be

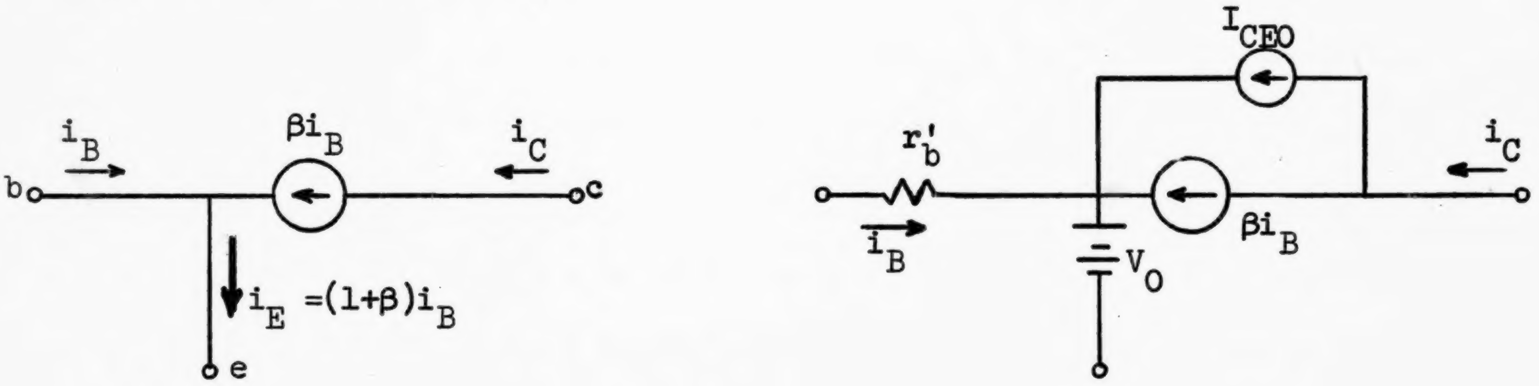
$$i_C = \frac{\alpha}{1-\alpha} i_B = \beta i_B \qquad (9\text{-}22)$$

where

$$\beta = \frac{\alpha}{1-\alpha} \qquad (9\text{-}23)$$

is the base-to-collector amplification factor. The closer α is to unity, the greater the value of β. (For $\alpha = .98$, $\beta = 49$).

Indeed, Eq. (9-22) is an analytical expression representing the family of collector characteristic curves. It gives a family of horizontal lines, on the i_C-v_{CE} axes, which are equally spaced for equal changes in i_B. It is thus an idealized characteristic. It is now possible to draw a model representing this behavior of a transistor, as shown in Fig. 9-13a.



(a) Fig. 9-13 (b)

The current source whose value is βi_B is a dependent source (also called a controlled source). Such a source is different from the kind of current source we have considered up till now which, by contrast, can be called an independent source. The value of the source current in Fig. 9-13a is dependent on the current somewhere else. Such dependent sources are characteristic of the models for the transistor and other electronic devices.

The circuit of Fig. 9-13a represents only the most important property of the transistor, its current amplification. It does not, however, account for some other effects. Sometimes it is important to take such secondary effects into account. The model in Fig. 9-13b gives a better approximation to the behavior of the transistor. The battery V_0 accounts for the base-to-emitter voltage which is typically less than one volt. (About .2 volt for germanium and .5 volt for silicon transistors.) The resistor r_b', called the base spreading resistance, accounts for the small voltage from the base terminal of the transistor to the interior of the base region. (A typical value of r_b' is 100 ohms.) And, finally, the current source labeled I_{CEO} is the collector current when the base current is zero. From Fig. 9-13b, the collector current has the value

$$i_C = \beta i_B + I_{CEO} \tag{9-24}$$

The effect of I_{CEO} is to cause the family of collector characteristics to start with a non-zero value of i_C when $i_B = 0$. I_{CEO} is quite sensitive to variations in temperature. It is typically zero at 25°C but becomes 10 ma. at 175°C (for a transistor whose maximum collector current is 30 ma.).

Small-signal Models

Up to this point, the transistor currents and voltages that have been under discussion have been the total instantaneous values. These total values have contributions from two sources, (1) from the bias sources, which serve to fix the quiescent operating point, and (2) from signals applied at the input. Once the quiescent operating point of the device has been fixed, we are subsequently interested in the variations of output that take place as a consequence of the signal. It would, therefore, be convenient to deal only with these incremental changes from the average values; for example, with i_{ce} and v_{ce} rather than i_{CE} and v_{CE}.

The model of a transistor which we shall now develop is obtained by an approach based on empirically obtained measurements of the terminal characteristics of a transistor, rather than from a study of the physical principles of semiconductor junctions.

To start the discussion, note that of the four terminal variables of the transistor (v_{BE}, i_B, v_{CE}, i_C), two can be expressed as functions of the other two. Look at the base and collector characteristics in Fig. 9-5. If we take i_B and v_{CE} as independent variables, then we can write

$$v_{BE} = f_1(i_B, v_{CE}) \tag{9-25}$$

$$i_C = f_2(i_B, v_{CE}) \tag{9-26}$$

where f_1 and f_2 are two functions of i_B and v_{CE} which are represented graphically by the curves in Fig. 9-5. Now suppose that both i_B and v_{CE} take on increments from their operating values. Then both v_{BE} and i_C will take on increments from their respective operating values. To a first approximation, the increments in the dependent variables will be linearly related to those of the independent variables. Thus, we can write

$$\left.\begin{aligned} \Delta v_{BE} &= a_1 \Delta i_B + a_2 \Delta v_{CE} \qquad (a) \\ \Delta i_C &= a_3 \Delta i_B + a_4 \Delta v_{CE} \qquad (b) \end{aligned}\right\} \quad (9\text{-}27)^*$$

---

[*] These can be obtained more rigorously by expanding the functions in Eqs. (9-25) and (9-26) in Taylor series about the operating point. Thus

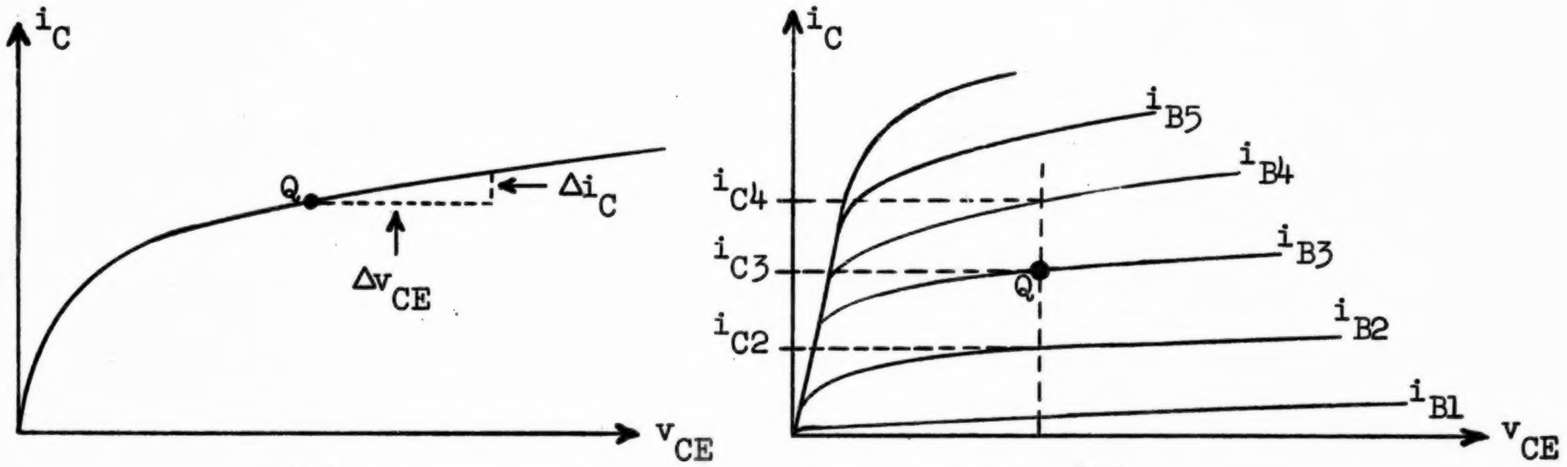
$$v_{BE} = v_{BE}\Big|_{\substack{\text{at operating}\\ \text{point}}} + \frac{\partial v_{BE}}{\partial i_B}\, di_B + \frac{\partial v_{BE}}{\partial v_{CE}}\, dv_{CE} + \begin{Bmatrix}\text{terms containing}\\ \text{higher derivatives}\end{Bmatrix}$$

The partial derivatives are to be evaluated at the operating point. If higher order terms are neglected, and we note that $v_{BE} - v_{BE}\Big|_{\text{at operating point}}$ is simply the increment in v_{BE}, the result follows.

The coefficients a_1 to a_4 are simply constants. Interpretations for these constants can be obtained by solving for them from the equations themselves. Thus,

$$
\begin{aligned}
a_1 &= \frac{\Delta v_{BE}}{\Delta i_B} \qquad \text{when } \Delta v_{CE} = 0 \qquad &\text{(a)}\\
a_2 &= \frac{\Delta v_{BE}}{\Delta v_{CE}} \qquad \text{when } \Delta i_B = 0 \qquad &\text{(b)}\\
a_3 &= \frac{\Delta i_C}{\Delta i_B} \qquad \text{when } \Delta v_{CE} = 0 \qquad &\text{(c)}\\
a_4 &= \frac{\Delta i_C}{\Delta v_{CE}} \qquad \text{when } \Delta i_B = 0 \qquad &\text{(d)}
\end{aligned}
\qquad (9\text{-}28)
$$

Further interpretations are obtained graphically from the transistor characteristic curves. Consider Eq. (9-28d) for a_4. Holding the increment of i_B at zero means the value of i_B stays constant. Thus, let Fig. 9-14a be a collector characteristic curve for a fixed value of i_B and let the coordinates of Q be the operating values of i_C and v_{CE}. We are to hold i_B constant at this value while i_C and v_{CE} change. The ratio of Δi_C to Δv_{CE} is simply the slope of the curve, provided the curve is approximately linear.



(a) (b)

Fig. 9-14

But even if there is some curvature in the curve, this ratio will approximately equal the tangent to the curve at the point Q, provided the increments Δi_C and Δv_{CE} are "small".

Now look at the interpretation of a_3 in Eq. (9-28c). Let Q in Fig. 9-14b be the operating point. $\Delta v_{CE} = 0$ means we are to hold the value of v_{CE} constant. Thus, however i_B and i_C change, the new operating point must lie on the vertical line passing through Q. Thus, if i_B changes from i_{B3} to i_{B4}, then i_C will change from i_{C3} to i_{C4}; so, a_3 will be the ratio of $i_{C4} - i_{C3}$ to $i_{B4} - i_{B3}$. Similar interpretations can be given for a_1 and a_2.

Glance back at Eqs. (9-27 or (9-28). The a's do not all have the same dimensions. Thus, a_3 is dimensionless and a_4 is a conductance. It is therefore more appropriate to give them more suggestive symbols. Earlier, we had introduced a notational scheme for expression varying components of the transistor currents and voltages. We will use this scheme and rewrite Eqs. (9-27) as follows.

$$v_{be} = r_{11}i_b + \mu v_{ce} \qquad \text{(a)}$$

$$(9\text{-}29)$$

$$i_c = \beta i_b + g_{22}v_{ce}$$

Remember that the voltage and current symbols here stand for changes, variations or increments in the instantaneous values. Thus, i_b is not the base current but the incremental base current.

We now have a set of equations; the next task is to draw a model which these equations imply. The first equation is suggestive of Kirchhoff's voltage law; the voltage v_{be} equals the sum of two voltages. One of these two is the drop across a resistor r_{11}; the other appears to be a dependent source. A similar analysis can be made of the second equation. Remembering that $i_e = i_c + i_b$, the result is the network shown in Fig. 9-15.

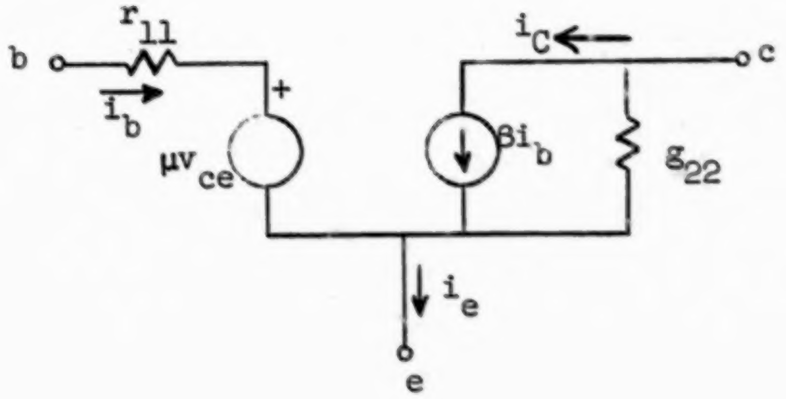


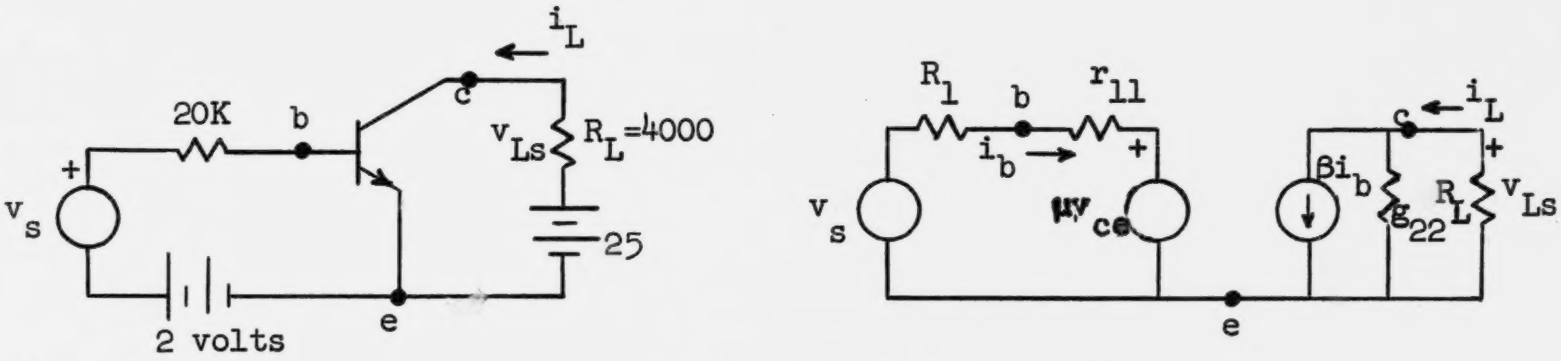
Fig. 9-15

This network is a model of a transistor. Remember the conditions under which it was derived. It does not give relationships among the total currents and voltages but only among changes in these currents and voltages around their operating values. The model includes two dependent sources; one a current source, the other a voltage source. Note that if the collector terminal is shorted to the emitter, then i_c equals βi_b and so β is the short circuit incremental current gain. It has the same value as the previously introduced β. Note also that no bias batteries appear in this incremental model.

On comparing this small-signal model with the model shown in Fig. 9-13, one can note that if r_{11}, μ, and g_{22} are all zero, the small-signal model reduces to the model in Fig. 9-13a. Thus, the basic nature of a transistor is a current amplifier; the other parameters simply measure the extent to which the transistor deviates from this ideal behavior. Because the parameters of this model are dimensionally so heterogeneous, it is called the hybrid model and the four parameters are called the hybrid parameters.

This model of a transistor can be used in the calculation of the gain of an amplifier. Figure 9-16 shows a typical transistor amplifier. It is desired to determine the small-signal voltage gain v_{Ls}/v_s, where v_s is the signal voltage and

v_{Ls} is the component of the voltage across R_L due to the signal. It is also desired to estimate the error in i_b that would result if μ were neglected. The parameters of the model are $r_{11} = 1$ kilohm, $\beta = 60$, $\mu = 4 \times 10^{-4}$, $g_{22} = 25$ micromhos; also $R_L = 4$ kilohms.



(a) (b)

Fig. 9-16

The first step is to draw the small-signal equivalent circuit. This is done by replacing the transistor with its small-signal model, removing all biasing sources and leaving everything else as is. The result is the diagram of Fig. 9-16. Let R be the parallel combination of the two resistors R_L and $1/g_{22}$. That is,

$$R = \frac{R_L}{1 + g_{22}R_L} = \frac{4000}{1 + 4000 \times 25 \times 10^{-6}} = \frac{4000}{1.1} = 3.64 \text{ kilohms.}$$

The incremental load voltage is

$$v_{Ls} = v_{ce} = -\beta i_b R = -.218 \times 10^6 \, i_b$$

Now this can be substituted for the dependent voltage source in the base circuit. The voltage of this source becomes $-\mu\beta R i_b$. This is the same voltage that would be obtained if a current i_b were flowing through a negative resistance $-\mu\beta R$. Thus, the base circuit can be redrawn as shown in Fig. 9-17. The negative resistance $-\mu\beta R$ subtracts from $R_1 + r_{11}$.

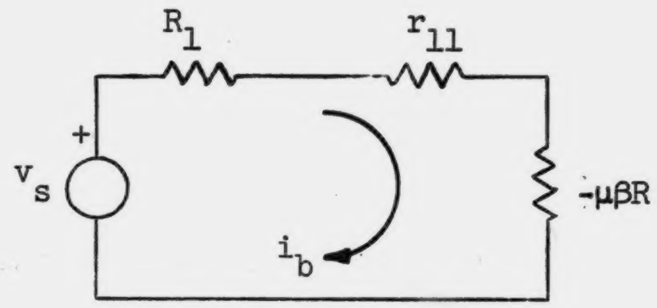


Fig. 9-17

The incremental base current can now be found to be

$$i_b = \frac{v_s}{R_1 + r_{11} - \mu\beta R} = \frac{v_s}{21000 - 87.2} = \frac{v_s}{21000} \frac{1}{1 - \frac{87.2}{21000}}$$

$$= \frac{v_s}{21000} \frac{1}{1 - .00415} = 1.004 \frac{v_s}{21000} \text{ amp.}$$

Note that, if μ were neglected and assumed to be zero, i_b would be simply $v_s/21000$. Thus, the error in neglecting μ would be only about .4 per cent, which is certainly tolerable. (In fact, actual resistors used for R_1 and R_L would have tolerances of 5 to 20 per cent, so that the additional .4 per cent error would be completely negligible.)

To complete the problem, the value of i_b just found is substituted into the previously found expression for load voltage to yield, for the voltage gain,

$$\frac{v_{Ls}}{v_s} = \frac{-\beta R}{R_1 + r_{11} - \mu\beta R} \doteq \frac{-\beta R}{R_1 + r_{11}} = -10.4 \qquad (9\text{-}30)$$

Note that the gain is a negative number. Thus, if the signal is sinusoidal, the output signal voltage will have a phase angle of 180° relative to the input signal, and its amplitude will be 10.4 times larger.

In this example, all the transistor and circuit parameter values were given and the gain was calculated. However, the expression for the gain can be examined to see what changes might be made to increase the gain. Of the four quantities on the right side of Eq. (9-30), β and r_{11} are fixed by the transistor. (Their values also depend on the location of the quiescent operating point; thus β depends on the separation of the curves in the family of collector characteristic curves and a glance at Fig. 9-5 shows that this separation depends on where the quiescent point is located. However, if we assume that the quiescent point remains somewhere near the center of the quadrant, the variations of the transistor parameters due to changes in the operating point -- which are determined by the external circuit parameters -- will be secondary. To a first approximation we can claim that β and r_{11} depend only on the transistor.) R_1 is an external resistor and R is a combination of an external resistor and the transistor parameter g_{22}. However, for the values in this example, the difference between R and R_L is less than ten per cent. Thus, R can be looked upon as an external parameter.

For a given transistor, then, in order to increase the amplifier gain we should increase R (which means increase R_L) or decrease R_1, or both. But note that R_L and R_1 help to establish the load line and fix the quiescent point, so we are not at liberty to change these values at will. A compromise must be reached between desirable quiescent point and large gain.

One additional variable available for achieving desired objectives is the transistor itself. It would be desirable to have transistors with large values of β and small values of r_{11}. The engineer wanting to optimize a circuit must have knowledge of the characteristics of a wide range of transistors; this is a by-produce of experience. With the introduction provided here, you can now interpret many transistor characteristic curves, you can design a variety of simple amplifiers, and you can begin to acquire this valuable engineering experience.

Practical Considerations in Amplifier Operation

In the preceding work, consideration was given only to the most basic amplifier configuration. This included two bias sources, one for the collector and one for the base, and two resistors, R_L in the collector circuit and R_1 in the base circuit. Collectively, these quantities served to establish the load line and fix the operating point on the load line.

One hallmark of an engineer is to see if he can accomplish a given task more simply, with fewer components. One candidate for such simplification is the use of two batteries for biasing purposes; can only one battery do the job? The question is answered in Fig. 9-18. The diagram in (a) shows an arrangement which uses only one battery but which is equivalent to the arrangement in (b). (If you don't see this quickly, look back at the discussion of the shifting of sources in Sec. 2-6.) The only difficulty here is that we are constrained to have the same value of biasing source in both collector and base circuits. We need to have some way of adjusting the value of the base biasing battery voltage independently from that of the collector biasing battery.

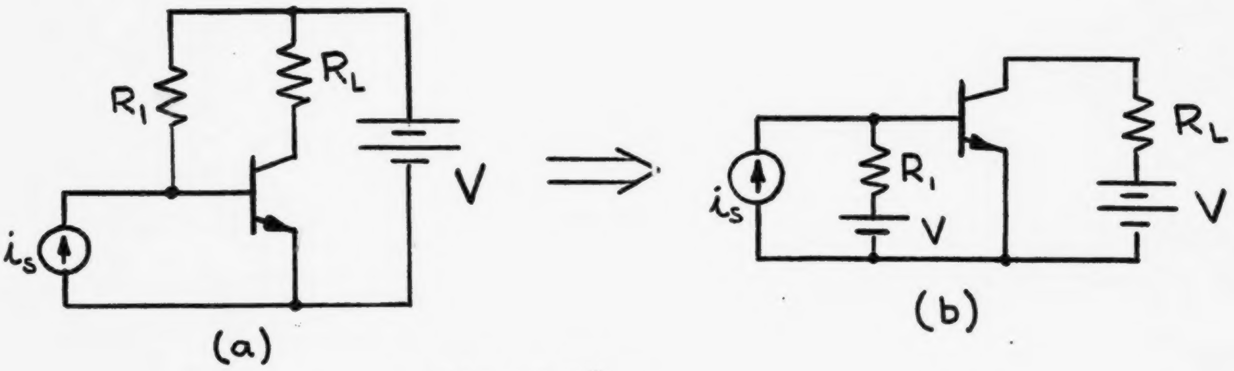


Fig. 9-18

This objective can be achieved by the arrangement shown in Fig. 9-19a. An extra resistor has been added in the base circuit forming a voltage-divider arrangement. If the structure to the left of the dashed vertical line is replaced by a

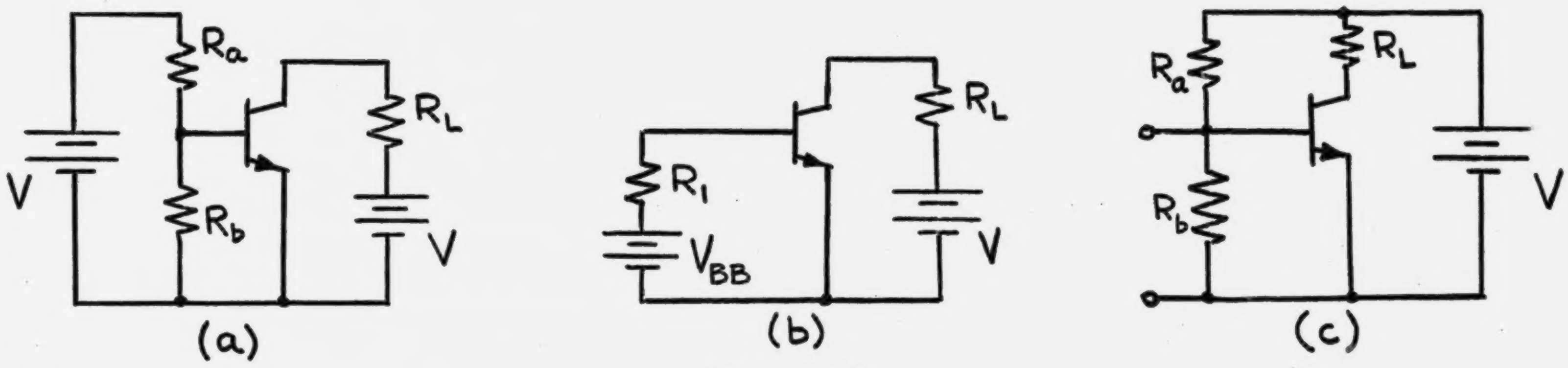


Fig. 9-19

Thevenin equivalent the result is as shown in Fig. 9-19b, where

$$V_{BB} = \frac{R_b}{R_a + R_b} V \qquad (a)$$

$$R_1 = \frac{R_a R_b}{R_a + R_b} \qquad (b) \qquad (9\text{-}31)$$

The final amplifier diagram has the arrangement shown in Fig. 9-19c. Note that the original two batteries and two resistors have been replaced by one battery and three resistors.

In a given problem if the component values are given and it is required to determine the quiescent operating point and the load line, Eqs. (9-31) are used to fix the values in the configuration of Fig. 9-19b. The rest of the procedure is then the same as previously described. On the other hand, if the problem is one of design, and the desired operating characteristics are given -- such as largest signal amplitude with least distortion, which approximately fixes the Q point and the load line -- the problem being to determine appropriate component values, one has to work backwards. Thus, for Fig. 9-19b, the base current can be found, using Eqs. (9-31), to be

$$i_B = \frac{V_{BB}}{R_1} = \frac{V}{R_a}$$

Thus, with V determined from the load line and i_B fixed by the quiescent point, R_a becomes known. There still remains to determine R_b from Eqs. (9-31). Since neither V_{BB} nor R_1 in Fig. 9-19b, nor R_b in Fig. 9-19c, have been fixed by anything specified to this point, only practical considerations and convenience will limit you in selecting their values. For example, you would not want R_b to be either very large (such as several million ohms) or very small (such as only a few ohms) as resistors in these sizes are expensive and fragile. Thus, if you are designing a circuit like Fig. 9-19c, you are free to choose any convenient value of R_b. However, you should be forewarned that other modifications in the amplifier circuit will be introduced, below. They will put further limits on you in selecting your circuit parameters.

At this point, let us use the model of Fig. 9-13 in the amplifier circuit of Fig. 9-19b to find an analytical expression for the dc collector current in order to investigate the factors on which it depends. The resulting network, which applies for the average values I_B and I_C (remember the notation), is shown in Fig. 9-20. By writing a voltage equation around the left-hand loop, I_B can be found. Then,

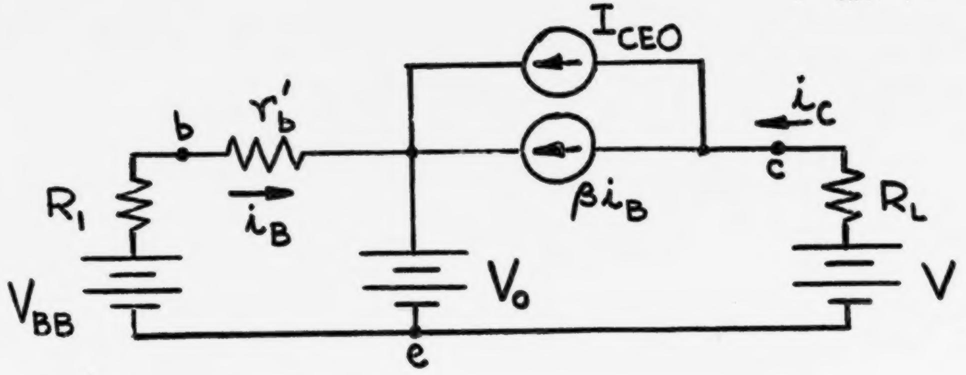


Fig. 9-20

a node equation at the upper right node gives an expression for I_C. Thus,

$$I_B(R_1+r'_b) = V_{bb} - V_0$$

$$I_C = \beta I_B + I_{CEO} \tag{9-32}$$

$$I_C = \frac{\beta(V_{BB}-V_0)}{R_1 + r'_b} + I_{CEO}$$

In the last expression, the quantities that are transistor parameters are β, V_0, I_{CEO} and r_b'. It has already been pointed out that I_{CEO} is temperature dependent. The same is true of V_0. Thus, I_C and the operating point will change with temperature. Furthermore, the value of β for a given type of transistor varies from one individual transistor to another, simply because of manufacturing problems, and this variation can be by a factor of as much as 3 or 4. Since β is a multiplying factor in one term in Eq. (9-32), I_C will change radically with such changes in β. A modification in the amplifier circuit is needed that will tend to compensate for changes in transistor parameters such as β.

A review of the steps in arriving at Eq. (9-32) shows that, although there are two loops in Fig. 9-20, the base current is found from the left-hand loop independently of the collector current. Thus, any changes in the collector current due to changes in β, say, are not reflected back to influence the base current -- we say there is no "coupling" between the two loops. This would not be the case if there were a resistor in the branch between two loops; that is, if a resistor were connected to the emitter in Fig. 9-19. Such a resistor is added in the amplifier circuit and its model shown in Fig. 9-21. This resistor tends to compensate for changes in β.

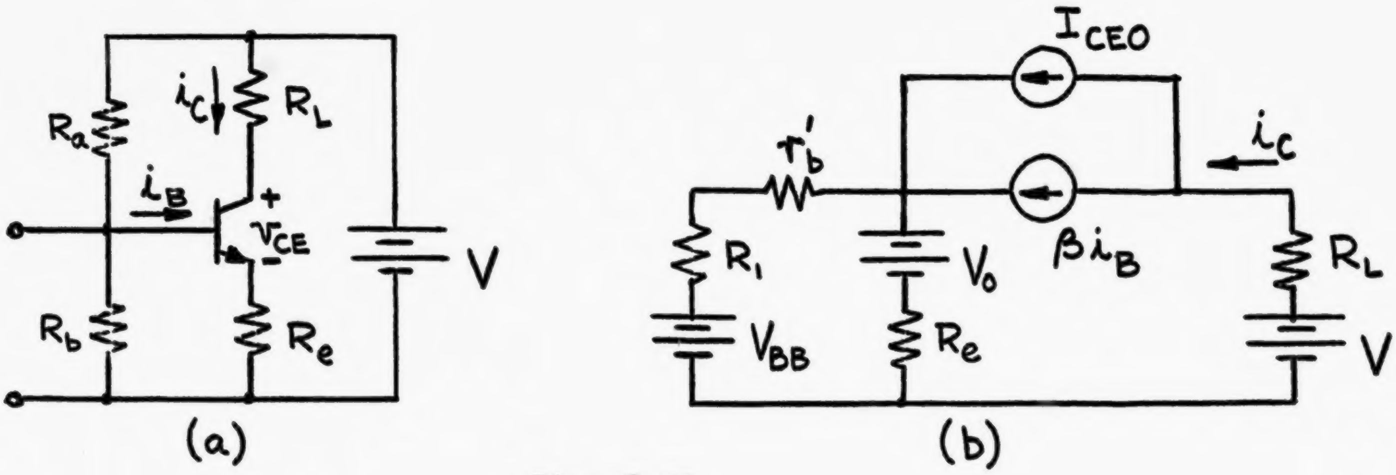


Fig. 9-21

With this change, the equations written earlier now become

$$(r_b' + R_1 + R_e)I_B + R_e I_C = V_{BB} - V_0 \qquad \text{(a)}$$

$$I_C = \beta I_B + I_{CEO} \qquad \text{(b)} \qquad (9\text{-}33)$$

When these are solved for I_B and I_C, the result will be

$$I_B = \frac{V_{BB} - V_0 - R_e I_{CEO}}{r_b' + R_1 + (\beta+1)R_e} \qquad (9\text{-}34)$$

$$I_C = \frac{\beta(V_{BB} - V_0)}{r_b' + R_1 + (\beta+1)R_e} + \frac{r_b' + R_1 + R_e}{r_b' + R_1 + (\beta+1)R_e} I_{CEO} \qquad (9\text{-}35)$$

(Verify that these reduce to the previous expressions when $R_e = 0$.) Consider the first of these for I_B. Suppose there is a positive increment in either I_{CEO} or β. An increase in β increases the denominator and thus causes a decrease of I_B. Similarly, an increase in I_{CEO} causes a decrease of the numerator and again a decrease of I_B. Thus, an increase in either β or I_{CEO}, or both, tends to reduce I_B. Considering Eq. (9-33b), this decrease in I_B tends to compensate for the increase in β or I_{CEO}, thus tending to keep I_C constant. The operation of the amplifier will thus be greatly improved by the addition of R_e.

Now concentrate on the last equation for I_C (Eq. 9-35). This appears to be a complicated expression, although it is merely the sum of two terms. It is to be compared with Eq. (9-32) for the case of no R_e. In that expression the coefficient of I_{CEO} is 1. Let us examine what value of R_e would be required to make this coefficient in Eq. (9-35) considerably less than 1, for typical values of the other quantities.

Let us call the coefficient of I_{CEO}, A, and solve for R_e. The result will be

$$\frac{r_b' + R_1 + R_e}{r_b' + R_1 + (\beta+1)R_e} = A$$

or

$$R_e = \frac{(1-A)(r_b'+R_1)}{A(\beta+1) - 1} \tag{9-36}$$

Typical values of β, r_b' and R_1 are 50, 100 ohms, and 5-10 kilohms, respectively. Thus, r_b' is quite negligible compared to R_1. If we wish the coefficient A to be 0.1, then

$$R_e = 1.1 \text{ to } 2.2 \text{ kilohms for } R_1 = 5 \text{ to } 10 \text{ kilohms.}$$

These are quite reasonable values of R_e. Thus, the coefficient of I_{CEO} in Eq. (9-35) can be easily made of the order of .1. This means, the contribution of I_{CEO} to I_C can be made small enough to neglect. Thus, Eq. (9-35) reduces to

$$I_C = \frac{\beta(V_{BB}-V_0)}{r_b' + R_1 + (\beta+1)R_e} \tag{9-37}$$

This equation is quite useful in determining the collector current at the quiescent point. Of course, this point must also fall on the load line. The equation for the load line, however, will now be somewhat more complicated due to the appearance of the emitter resistor R_e. Thus, from the collector circuit in Fig. 9-21a, the following equation results.

$$v_{CE} = V - (R_L+R_e)i_C - R_e i_B \tag{9-38}$$

Not only does R_e contribute to the coefficient of i_C, it introduces i_B into this equation.

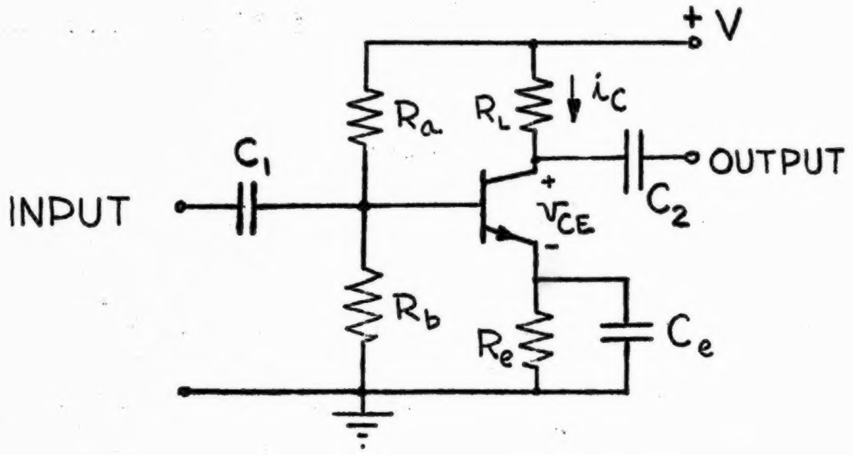
Of course, i_B could be eliminated by writing another equation for the base circuit. However, it would not cause a great error if i_B is neglected completely, since it is quite small compared with i_C, assuming a large enough β. If i_B is neglected, Eq. (9-38) can be solved for i_C to yield

$$i_C = \frac{V}{R_L + R_e} - \frac{1}{R_L + R_e} v_{CE} = \frac{V - v_{CE}}{R_L + R_e} \qquad (9\text{-}39)$$

Thus, the slope of the load line is determined by the sum of R_L and R_e, rather than just R_L as in the simpler circuit.

The Dynamic Load Line

The introduction of an emitter resistance has been shown to stabilize the value of collector current against changes in the transistor parameters, β and I_{CEO}. The same mechanism that brings about this improvement also brings about a reduction in the amplification of signals. But, we can modify the circuit further, to prevent this reduction. As far as the signal components of collector current are concerned, we would like to remove R_e, yet we would like to keep it there as far as the dc component is concerned. This objective can be achieved by placing a capacitor in parallel with R_e. Such an arrangement is shown in Fig. 9-22. R_e is said to be by-passed and C_e is a by-pass capacitor.



(The purpose of C_1 and C_2 will be discussed in the next few paragraphs.)

Fig. 9-22

Before considering the influence of the capacitors, note that the battery does not appear explicitly. Instead, the terminal to which the positive terminal of the battery should be connected is labeled + V; it is understood that the other terminal of the battery is connected to ground, which is designated by the symbol ⏚ at the bottom of the diagram.

Now the average value of the current in a capacitor must be zero. Another way of saying this is that a capacitor cannot carry direct current. Hence, for calculation of quiescent values, capacitors behave like open circuits and can simply be removed from the circuit and omitted from any calculations. This means, the presence of C_e in parallel with R_e will not influence the value of I_C, the average collector current, obtained in the last section.

Now let us consider the effect of the parallel combination of R_e and C_e on the signal components of current. Let the impedance of this combination be $\bar{Z}_e$ and assume the signal to be sinusoidal of angular frequency ω. Then

$$\bar{Z}_e = \frac{R_e \frac{1}{j\omega C_e}}{R_e + \frac{1}{j\omega C_e}} = \frac{R_e}{1 + j\omega C_e R_e} \qquad (9\text{-}40)$$

If $\omega C_e R_e$ is much greater than 1, this expression becomes approximately $1/j\omega C_e$ and the effect of R_e has been eliminated. For example, let R_e = 500 ohms and let the lowest frequency signal to be amplified have ω = 5000 rad/sec. What value of C_e will cause $\omega C_e R_e$ to equal, say, 25 and what will be the value of $\bar{Z}_e$ for this value of C_e? The answer will be

$$C_e = 25/\omega R_e = 10 \text{ mfd}; \text{ and } \bar{Z}_e = -\,j20$$

Now, 20 ohms is practically a short circuit when compared with the value of R_e = 500 ohms. At higher frequencies this same value of capacitance will make the value of $\bar{Z}_e$ even smaller and more like a short circuit. For the signal components, then, the capacitor can be taken as a short circuit, thus eliminating R_e and restoring the signal amplification that we had before adding R_e. Note that a by-pass capacitor must have a large capacitance, such as 10 mfd.

As for the capacitors labeled C_1 and C_2 at the input and output of the amplifier in Fig. 9-22, they also are effectively an open circuit for dc calculations. At the same time, if their values are made large enough, each will be effectively a short circuit at the signal frequencies. Thus, they have two functions: (1) to couple the time-varying part of the signal into the amplifier and out of it, and (2) to block out any average component it might have. (Why would it be undesirable for a signal with a nonzero average value to be applied to the amplifier?) Thus, these capacitors are called either of two descriptive names, coupling capacitors and also blocking capacitors.

The discussion concerning the effect of the by-pass capacitor has shown that C_e acts as a short circuit to the signal component of emitter current. Thus, the voltage across R_e is caused only by the average value of the emitter current, and therefore remains constant at the value $R_e I_C$. Thus, writing a voltage equation around the collector circuit in Fig. 9-22 and solving for i_C leads to

$$i_C R_L + v_{CE} + I_C R_e - V = 0$$

$$i_C = \left(\frac{V - R_e I_C}{R_L}\right) - \frac{1}{R_L} v_{CE} \qquad (9\text{-}41)$$

This is the equation of a straight line. It should be compared with the equation of the load line for the circuit without C_e, given in Eq. (9-29). Its slope is different, as are its intercepts. However, it passes through the same quiescent operating point, as can be confirmed by setting i_C in Eq. (9-41) equal to its quiescent value I_C, and noting that the resulting values of v_{CE} in the two equations are the same. This new line is called the dynamic load line in contrast with the previous one which is the static load line. Figure 9-23 shows both lines on a set of collector characteristic curves.

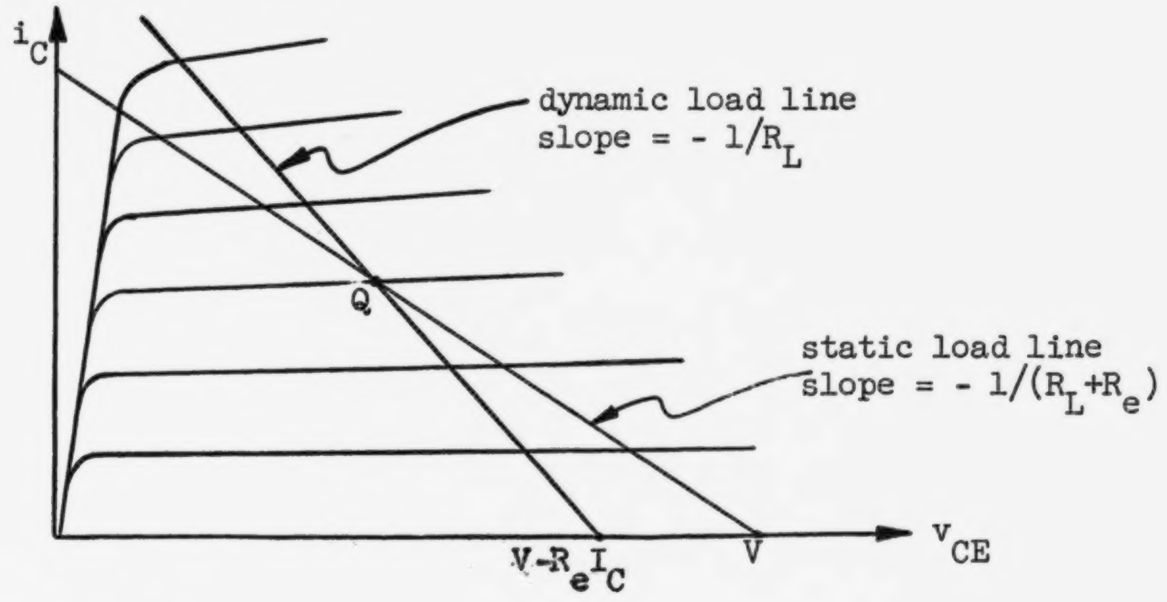

i_C
dynamic load line
slope = - 1/R_L
Q
static load line
slope = - 1/(R_L+R_e)
V-R_eI_C
V
v_CE

Fig. 9-23

Example

To illustrate the above discussion, consider the amplifier shown in Fig. 9-24.

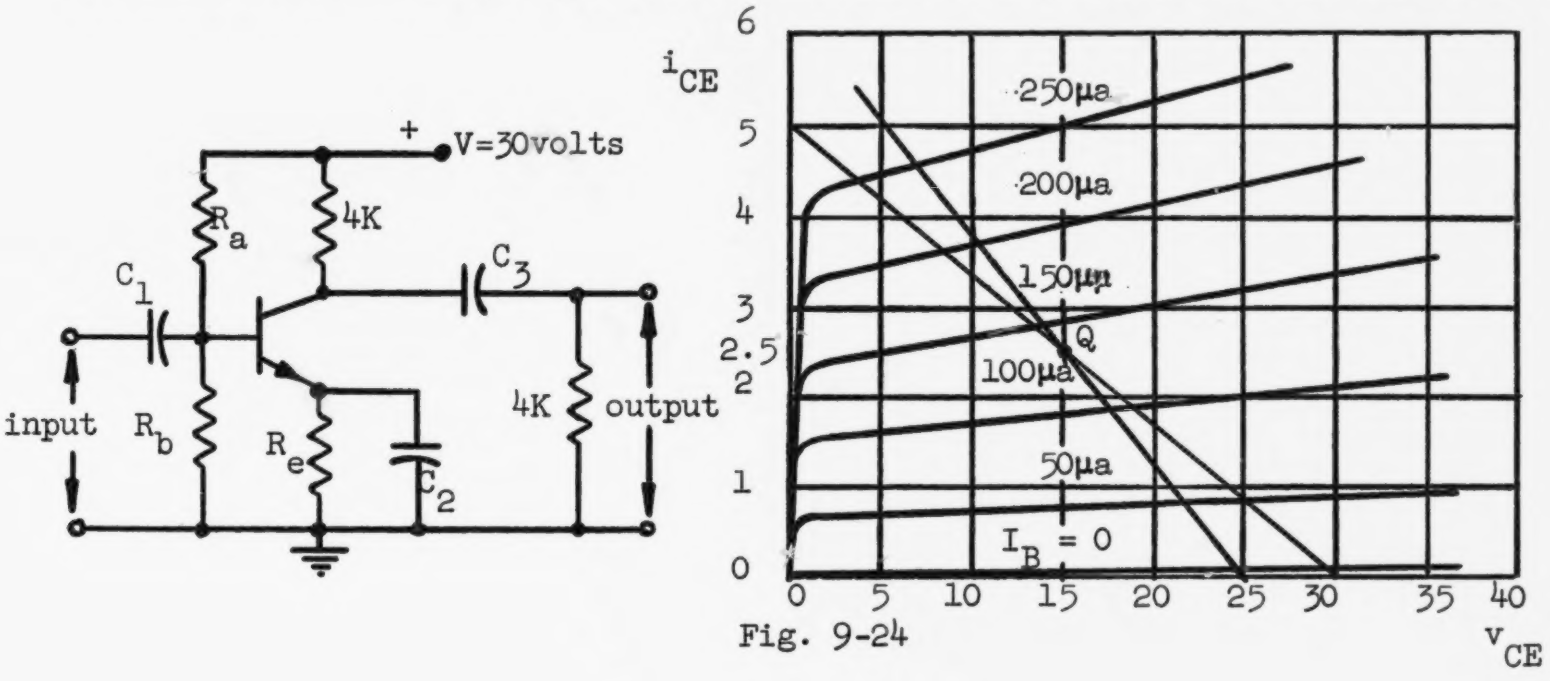

+
V=30volts
R_a
4K
C_1
C_3
input
R_b
R_e
C_2
4K
output
i_CE
6
5
4
3
2.5
2
1
0
250μa
200μa
150μa
Q
100μa
50μa
I_B = 0
0
5
10
15
20
25
30
35
40
v_CE

Fig. 9-24

The transistor has a V_0 = .2 volts. It is desired to have the quiescent operating point fall at i_C = 2.5 ma., v_{CE} = 15 volts. Draw the static and dynamic load lines and determine appropriate values for those parameters which are unspecified in the diagram. Use the collector characteristics shown in the figure.

The value of β in the vicinity of the Q point is found from the collector curves to be $\beta \doteq 20$. The static load line is easily constructed since it must pass through the v_{CE} axis at 30 volts and through the Q point. It intersects the i_C axis at 5 ma. Hence, the slope will be

$$\text{slope} = -\frac{5}{30}\,\frac{\text{ma}}{\text{volt}} = -\frac{1}{R_e + R_L}$$

Hence,

$$R_e + R_L = \frac{30}{5} = 6 \text{ kilohms}$$

Since $R_L = 4K$, then $R_e = 2K$.

The dynamic load line can now be drawn. Perhaps the simplest way is to note that it must pass through the Q point and through the point $V - R_eI_C = 30 - 2(2.5) =$ $= 25$ volts on the v_{CE} axis.

It remains to find R_a and R_b. Since the quiescent collector current is known, Eq. (9-37) gives a relationship between V_{BB} and R_1. Thus,

$$R_1 + 42 = \frac{20}{2.5}(V_{BB} - .2)$$

or

$$R_1 = 8V_{BB} - 43.6 \text{ kilohms.}$$

When Eqs. (9-31) are used for R_1 and V_{BB}, the result will become a relationship between R_a and R_b. Thus,

$$\frac{R_b(240 - R_a)}{R_a + R_b} = 43.6$$

From this expression it is clear that R_a must be less than 240K. There are no other restrictions on R_a and R_b. Hence, let us arbitrarily choose $R_a = 150K$. Then, from the last expression we find $R_b = 141K$.

Thus we have "designed" a transistor amplifier with predetermined characteristics. It would still be necessary to build this design and try it out in the laboratory. Some of the values may need to be adjusted slightly in order to get just the desired results, but your analysis has done two things: (1) you have saved a lot of time and money in getting a design that should work, and (2) you will understand the effect of making small adjustments in the values of the different components, to get just the performance you want.

Let us continue with this example and now concentrate on finding the small signal gain of the amplifier assuming a signal current i_1 (milliamps) is applied. For this purpose, remember that the capacitors C_1, C_2 and C_3 are assumed to be short circuits and biasing batteries are removed (assumed shorted). Thus, the resulting network is shown in Fig. 9-25a. Note that the 4K output resistor and R_L are now in parallel since C_3 is a short circuit, and R_e has been shorted by C_2. (R_a and R_b are also in parallel to yield R_1.) The transistor is now replaced by its small signal model from Fig. 9-15 to yield the result shown in Fig. 9-25b. It has here been assumed that μ is small enough to neglect. β has already been found to be $\beta \doteq 20$. From Fig. 9-24, g_{22} is found to be $g_{22} \doteq .03$ mmho. Although there is no data from which r_{11} can be determined, it is clear that if it has a typical value in the neighborhood of 1K, it will be overshadowed by R_1. Let us assume r_{11} = .4K in the interest of convenience and convert i_1 in parallel with R_1 to a voltage

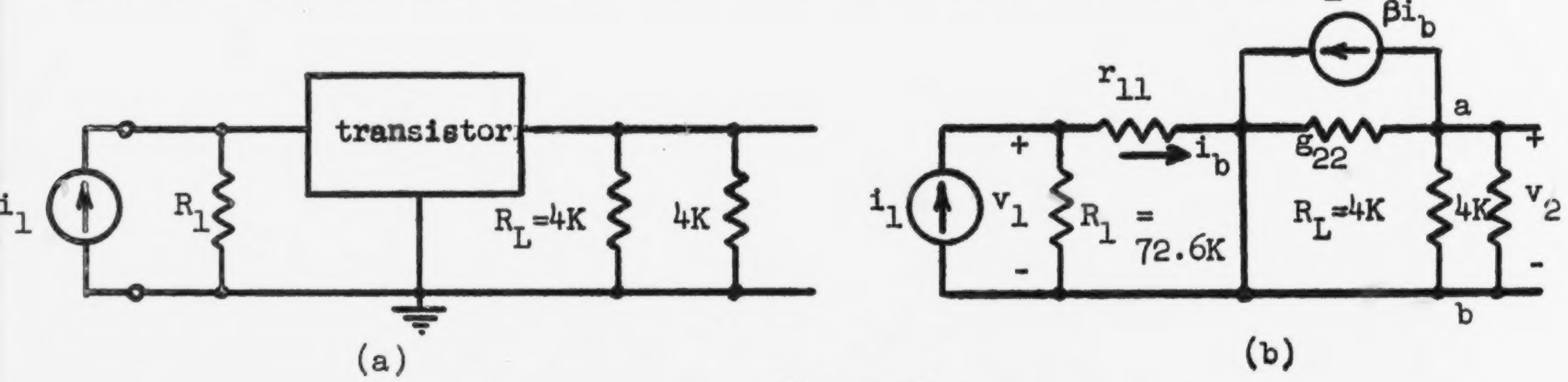


Fig. 9-25

source equivalent and similarly with βi_b in parallel with g_{22}. (Caution: g_{22} is a conductance.) The result is the network shown in Fig. 9-26.

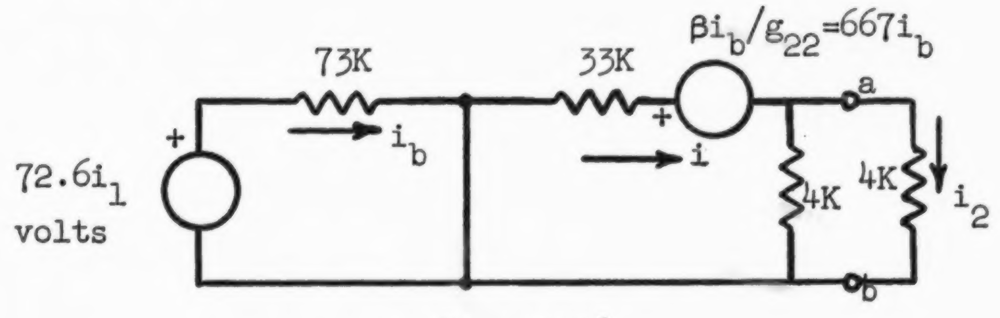


Fig. 9-26

This is relatively easy to solve because each of the two loops, considering the two 4K resistors to be combined, can be handled separately.

$$i_b = \frac{72.6}{73} i_1 \doteq i_1 \text{ ma.}$$

$$i = \frac{-667i_b}{33 + 2} = -19i_1$$

Finally, the output current i_2 can be found from the current divider relationship to be

$$i_2 = \frac{4}{4+4} i = -9.5i_1 \text{ ma.}$$

giving a current gain of

$$\frac{i_2}{i_1} = -9.5$$

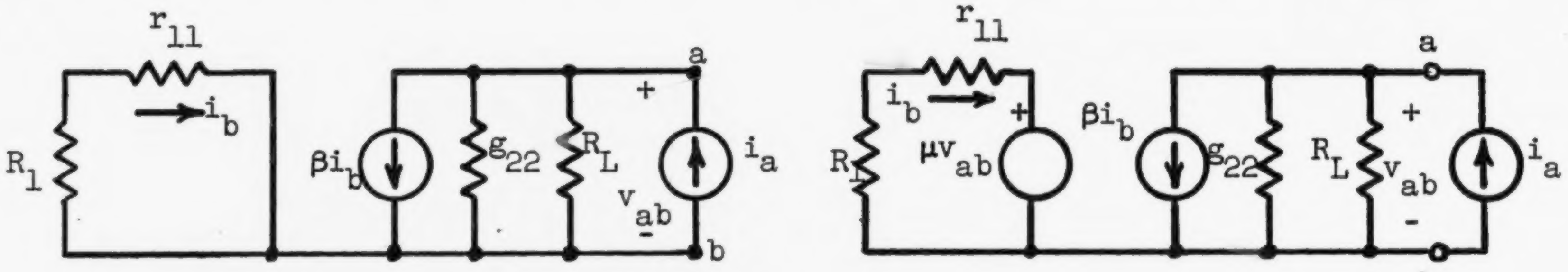
In addition to the gain, there are other quantities of interest in the small-signal operation of transistor amplifiers. One of these is the impedance at the input terminals. You will recall that the impedance of a network at a pair of terminals is the ratio of the phasor voltage at those terminals to the phasor current. This assumes that the signals applied to the amplifier are sinusoids. If the network in question has no reactive elements (capacitors or inductors), the angle of all sinusoids (currents or voltages) will be the same; there will be no phase difference, only the amplitudes will be different for different variables. In such a case, all phasors can be made real. Hence, it will be permissible to deal with the sinusoids themselves instead of their phasors. The input impedance is purely real and has no imaginary component.

Thus, for the amplifier network in Fig. 9-25b, the input impedance at the terminals of the amplifier will be v_1/i_1. From the figure it is seen that r_{11} and R_1 are in parallel and their voltage and current are not influenced by the collector circuit, because μ was assumed negligible. Thus, the input impedance is the resistance which is the equivalent of r_{11} and R_1 in parallel. This will be approximately equal to r_{11} since R_1 is so much larger than r_{11}. But r_{11} is quite small (we took it to be 400 ohms). This is typical of transistor amplifiers in this configuration.

Another quantity of interest is related to the Thévenin equivalent of the amplifier at its output terminals. Consider forming the Thévenin equivalent looking to the left from terminals a-b in either Fig. 9-25b or Fig. 9-26. Interest will be focussed on the Thévenin equivalent impedance. You will recall that this is the impedance at the terminals when the sources in the network are deactivated. Since we now have a type of source -- a dependent source -- which was not introduced when the Thévenin theorem was discussed, the question arises as to whether or not dependent sources are also deactivated when the Thévenin impedance is calculated. Note that in Fig. 9-25b the dependent source current is proportional to i_b. If this source is deactivated, βi_b, and hence also i_b, must be set equal to zero. But i_b is the current in r_{11}; it may or may not be zero. If i_b should turn out to be zero, then the dependent source current will be zero; otherwise not.

In the present case, let us deactivate the signal current i_1 (make it zero) since it is an independent source and then apply a current source i_a at the terminals a-b, our objective being to find v_{ab} so that the ratio v_{ab}/i_a can be calculated.

The situation is illustrated in Fig. 9-27a, where the network has been redrawn in



(a) Fig. 9-27 (b)

order to show more clearly that there is no coupling back from the collector circuit to the base circuit. Thus, even though a source i_a is applied, there will be no base current and hence βi_b will be zero. Hence, the Thévenin impedance will be the parallel combination of g_{22} and R_L. It is common terminology to call this Thévenin equivalent impedance, the output impedance. In this case, the output impedance will be $R_L/(1+g_{22}R_L) = 3.57K$.

Suppose that the parameter μ in the transistor model is not neglected. Then, Fig. 9-27a becomes modified as in Fig. 9-27b. Now there is coupling from the collector circuit to the base circuit. Hence, i_b will not be zero; it will be

$$i_b = -\frac{\mu v_{ab}}{R_1 + r_{11}}$$

From the collector circuit we find v_{ab} to be

$$v_{ab} = \frac{R_L}{1+g_{22}R_L}(i_a - \beta i_b) = \frac{R_L}{1+g_{22}R_L}\left(i_a + \frac{\mu\beta v_{ab}}{R_1 + r_{11}}\right)$$

The last step is obtained by substituting i_b from the preceding equation. Solving for v_{ab} leads to

$$\frac{v_{ab}}{i_a} = \frac{R_L}{1 + (g_{22} - \mu\beta)R_L} \qquad (9\text{-}42)$$

which reduces to the preceding expression for the output impedance when $\mu = 0$.

Although this example does not prove the general case, it is nevertheless a general result that when finding the Thévenin impedance of a network containing dependent sources (as well as independent ones), only the independent sources are deactivated -- not the dependent sources. We shall not prove this result here.

The Emitter Follower

The amplifier configuration we have been discussing is called a common emitter connection, because the emitter terminal is common between input and output. We have seen that this amplifier has a relatively low input impedance (several hundred ohms) and a relatively high output impedance (several kilohms). In many applications just the opposite is required: relatively high input impedance (several tens of thousands of ohms) and a relatively low output impedance (in the hundreds or less). We see from Eq. (9-42) that the output impedance depends strongly on R_L; it will become zero if $R_L = 0$. But in this case there will be no output voltage either, so nothing is gained by setting $R_L = 0$.

As another point of departure, note from Fig. 9-25b that the input impedance will increase if the resistance R_e is not bypassed so that it appears in the small signal model. These thoughts lead to consideration of the amplifier configuration shown in Fig. 9-28a.

The two modifications here are that R_L has been set equal to zero and R_e is not bypassed with a capacitor. Since there is no load resistance, the output must be taken somewhere else; it is taken across R_e. The resulting small-signal equivalent circuit is shown in Fig. 9-28b. This model is redrawn in a more convenient form in Fig. 9-28c, to bring out more clearly the output terminals. Note that R_e and g_{22} are in parallel; let us denote by R the parallel combination:

$$R = \frac{R_e}{1 + g_{22}R_e} \tag{9-43}$$

There are three quantities of interest in the operation of the emitter-follower: the current gain, the input impedance and the output impedance. To find the gain and the input impedance, suppose a current source i_s is applied, leading to Fig. 9-29. If v_2 is found, then the current through R_e will be simply

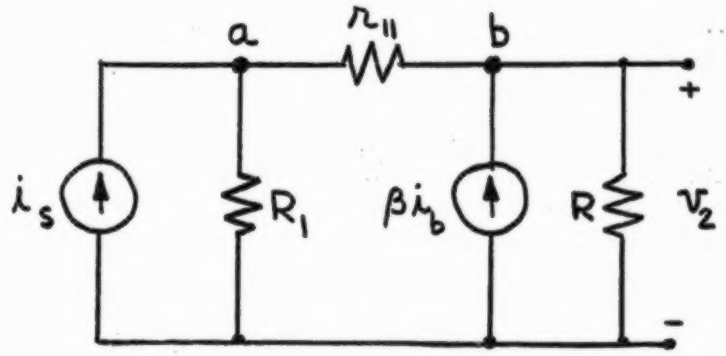


Fig. 9-29

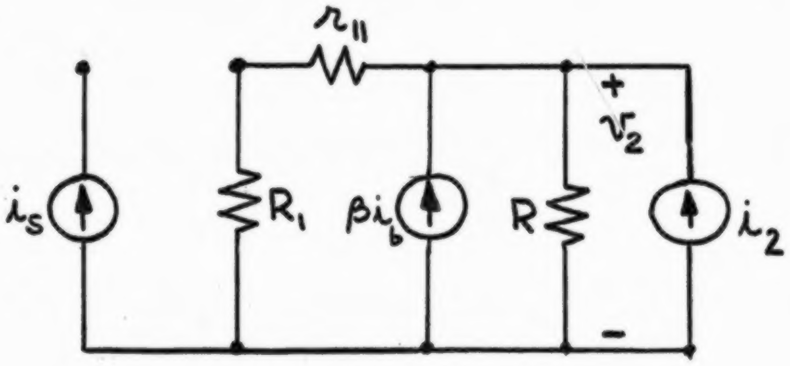


Fig. 9-30

v_2/R_e. Similarly, if v, is found, then the input impedance will be v_1/i_s. To find both of these voltages, one procedure is to write node equations. If this is done for the nodes labeled a and b in the diagram, the result will be

$$v_1\left(\frac{1}{R_1}+\frac{1}{r_{11}}\right)-\frac{1}{r_{11}}v_2=i_s$$

$$-\frac{1}{r_{11}}v_1+\left(\frac{1}{r_{11}}+\frac{1}{R}\right)v_2-\beta i_b=0 \qquad (9\text{-}44)$$

But i_b can be found in terms of v_1 and v_2 as

$$i_b=\frac{v_1-v_2}{r_{11}} \qquad (9\text{-}45)$$

This is now substituted in the previous equation and the node equations are solved for v, and v_2. The current gain and input impedance are found to be

$$\text{current gain} = G_i = \frac{\text{current in } R_e}{i_s} = \frac{(1 + \beta)R_1}{(R_1 + r_{11})(1 + g_{22}R_e) + (1 + \beta)R_e}$$

$$\doteq \frac{(1 + \beta)R_1}{R_1 + (1 + \beta)R_e} \qquad (9\text{-}46)$$

$$\text{input impedance} = Z_1 = \frac{v_1}{i_s} = \frac{R_1[r_{11}(1 + g_{22}R_e) + (1 + \beta) R_e]}{(r_{11} + R_1)(1 + g_{22}R_e) + (1 + \beta)R_e}$$

$$\doteq \frac{R_1 \, (1 + \beta)R_e}{R_1 + (1 + \beta)R_e} \qquad (9\text{-}47)$$

These expressions appear to be fairly complicated but they can be simplified as shown if all the resistors are assumed to have typical values. ($r_{11} \doteq 500$ ohms, $g_{22} \doteq 10^{-4}$ mhos, $\beta = 50$, R_1 in kilohms and $R_e \doteq 1$ K.) To get an impression of the values and how they compare with the previously discussed common emitter amplifier, let $R_1 = 72.6$K, $R_e = 2$K and $\beta = 50$ as in the example of the last section. Then,

$$G_i = 22.2$$

$$Z_i = 42.5\text{K}$$

It is noteworthy that the gain is about a factor of 2 greater than in the common emitter connection. Furthermore, it is positive, indicating that the input and output signals will be in phase.

As for the input impedance, the 42.5K is to be compared with the several hundred ohms represented by r_{11}.

The final quantity of interest is the output impedance. As discussed in the last section, to find this impedance only independent sources are deactinated. In Fig. 9-30, the signal source i_s has been open circuited and an external current source i_2 applied. It is required to find v_2. Applying Kcl at the upper node leads to

$$i_2 + \beta i_b = \frac{v_2}{R} + \frac{v_2}{R_1 + r_{11}} \tag{9-48}$$

Furthermore,

$$i_b = - \frac{v_2}{R_1 + r_{11}} \tag{9-49}$$

Substituting this into Eq (9-48) and rearranging leads to

$$\text{output impedance } Z_2 = \frac{v_2}{i_2} = R_1 + \frac{R(R_1 + r_{11})}{r_{11} + (\beta + 1)R}$$

$$= \frac{R_1 R_e}{R_1 + (\beta + 1)R_e} = \frac{Z_1}{1 + \beta} \tag{9-50}$$

Using the same numerical values as before, finally gives:

$$Z_2 = 830 \text{ ohms}$$

Equation (9-50) shows that large Z, and small Z_2 are conflicting requirements since Z_2 is proportional to Z_1. It is seen, however, that the larger the value of β, the better for achieving both objectives. Furthermore, a larger β also means a larger gain. Thus, one of the goals of transistor manufacturing is to make the β's as large as possible.

Chapter 10

Magnetic Coupling

Introduction

By virtue of the phenomena of magnetic flux and Faraday's law it is possible to transfer energy from one circuit to another, even though the circuits are not connected by wires. The practical device based on this fact is called a "transformer". In some cases, as in communication circuits, two coils are located near each other, in air, as in Fig. 10-1. In power applications, the coils are wound on an iron core, as in Fig. 10-2. The principles of analysis are essentially the same in both cases, with the exception that when there is an iron core it may be necessary to take non-linear effects into consideration.

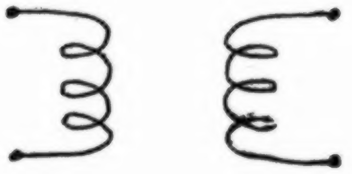


Figure 10-1.

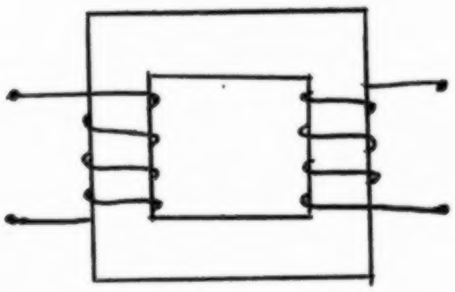


Figure 10-2.

10-1. Flux Relationships

Consider the pair of coils in Fig. 10-3, which are drawn in a specific way to facilitate the writing of equations. It is assumed that each coil is connected to a circuit which is not shown, so that currents i_1 and i_2 can flow. In general, these currents are varying with time.

Fluxes ϕ_1 and ϕ_2 respectively link coils 1 and 2. For the moment, assume there is no iron present, and that the resistance of each coil is negligible.

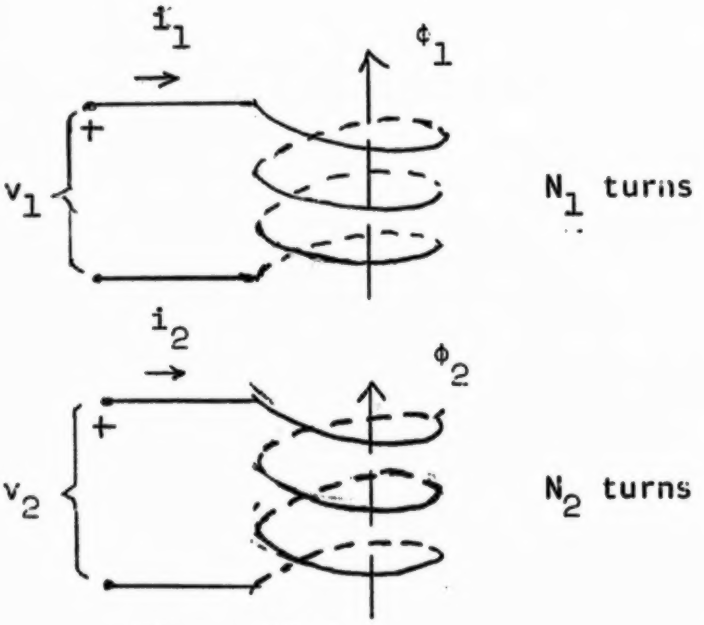


Figure 10-3.

There will be an emf induced in each coil which makes voltages v_1 and v_2 appear at the terminals. These are related to ϕ_1 and ϕ_2 by

$$v_1 = N_1 \frac{d\phi_1}{dt} \quad , \qquad v_2 = N_2 \frac{d\phi_2}{dt} \qquad (10\text{-}1)$$

Next we consider how ϕ_1 and ϕ_2 are related to i_1 and i_2. Without bothering with a specific case, which could in fact be quite complicated, it is possible to arrive at the required general relationship by the following arguments. If i_2 were zero, ϕ_1 would be proportional to i_1, and we could use a symbol k_{11} as the proportionality constant, to write

$$\phi_1 = k_{11} i_1$$

Furthermore, it would be possible that i_1 could be zero, and i_2 not zero. Due to proximity of the two coils, there could still be a flux ϕ_1, which would now be proportional to i_2. If k_{12} is the proportionality constant, this would be written

$$\phi_1 = k_{12} i_2$$

We have stipulated that no iron is present, so that linearity can be assumed, and therefore that superposition can be applied. Thus, with i_1 and i_2 each non-zero, we will have a general formula

$$\Phi_1 = k_{11}i_1 + k_{12}i_2 \tag{10-2}$$

A similar treatment can be applied to coil 2, giving

$$\Phi_2 = k_{21}i_1 + k_{22}i_2 \tag{10-3}$$

In doing this, four unknown constants $(k_{11}, k_{12}, k_{21}, k_{22})$ have been introduced. However, these will presently be eliminated, their purpose being merely to show how Φ_1 and Φ_2 vary with the two currents.

Now we can use the above equations in Eqs. (10-1) to give

$$\begin{aligned} v_1 &= k_{11}N_1 \frac{di_1}{dt} + k_{12}N_1 \frac{di_2}{dt} \\ v_2 &= k_{21}N_2 \frac{di_1}{dt} + k_{22}N_2 \frac{di_2}{dt} \end{aligned} \tag{10-4}$$

Again temporarily consider the special case where i_2 is identically zero (or constant) so that $di_2/dt = 0$. Then the first of the above equations becomes

$$v_1 = k_{11}N_1 \frac{di_1}{dt}$$

However, we would also have

$$v_1 = L_1 \frac{di_1}{dt}$$

where L_1 is the self inductance of coil 1. Therefore, $k_{11}N_1 = L_1$. Likewise, consideration of coil 2 gives $k_{22}N_2 = L_2$. Therefore, the equations can be written

$$\begin{aligned} v_1 &= L_1 \frac{di_1}{dt} + k_{12}N_1 \frac{di_2}{dt} \\ v_2 &= k_{21}N_2 \frac{di_1}{dt} + L_2 \frac{di_2}{dt} \end{aligned} \tag{10-5}$$

It can be shown that*

$$k_{12}N_1 = k_{21}N_2$$

Let this common value be represented by the letter M. We then have

$$M = k_{12}N_1 = \left(\frac{k_{12}}{k_{11}}\right)L_1$$

$$M = k_{21}N_2 = \left(\frac{k_{21}}{k_{22}}\right)L_2$$

or

$$M = \sqrt{\left(\frac{k_{12}}{k_{11}}\frac{k_{21}}{k_{22}}\right)}\sqrt{L_1L_2}$$

It is customary to use the symbol k to represent the first radical, namely

$$k = \sqrt{\frac{k_{12}}{k_{11}}\frac{k_{21}}{k_{22}}} \qquad (10\text{-}6)$$

It is called the *coefficient of coupling*, and we finally have

$$M = k\sqrt{L_1L_2} \qquad (10\text{-}7)$$

M is called the *mutual inductance* between the coils. k is always less than 1.

---

*Assume i_1 and i_2 are sinusoidal of frequency ω (certainly an acceptable special case). They will be represented by phasors $\bar{I}_1$ and $\bar{I}_2$. Recalling that differentiation produces a multiplication by $j\omega$ in phasor notation, we have

$$\bar{V}_1 = j\omega(L_1\bar{I}_1 + k_{12}N_1\bar{I}_2)$$

$$V_2 = j\omega(k_{21}N_2\bar{I}_1 + L_2\bar{I}_2)$$

Recall that power in the sinusoidal case is $\mathrm{Re}(\bar{V}\bar{I}^*)$. The powers into coils (1) and (2) are respectively

$$P_1 = \mathrm{Re}(\bar{V}_1\bar{I}_1^*) = \mathrm{Re}\left[j\omega(L_1\bar{I}_1\bar{I}_1^* + k_{12}N_1\bar{I}_2\bar{I}_1^*)\right]$$

$$P_2 = \mathrm{Re}(\bar{V}_2\bar{I}_2^*) = \mathrm{Re}\left[j\omega(k_{21}N_2\bar{I}_1\bar{I}_2^* + L_2\bar{I}_2\bar{I}_2^*)\right]$$

However, the product of a complex number by its conjugate is real. Hence $j\omega L_1\bar{I}_1\bar{I}_1^*$ and $j\omega L_2\bar{I}_2\bar{I}_2^*$ are each pure imaginaries and have zero real parts. Also, if $\bar{A}$ is a complex number, $\mathrm{Re}(j\bar{A}) = -\mathrm{Im}(\bar{A})$. Hence

$$P_1 = -\mathrm{Im}(\omega k_{12}N_1\bar{I}_2\bar{I}_1^*)$$

$$P_2 = -\mathrm{Im}(\omega k_{21}N_2\bar{I}_1\bar{I}_2^*) = \mathrm{Im}(\omega k_{21}N_2\bar{I}_2\bar{I}_1^*)$$

(continued on pg. 5)

Thus, it is seen that the original set of four k's have disappeared, having been replaced by three measurable parameters L_1, L_2, and k (or M). Note that four parameters reduced to three because one of the original four was not independent, since $k_{12}N_1 = k_{21}N_2$.

10-2. The General Equations

Having developed the concepts of mutual inductance, Eqs. (10-5) can now be written

$$v_1 = L_1 \frac{di_1}{dt} + M \frac{di_2}{dt}$$
$$v_2 = M \frac{di_1}{dt} + L_2 \frac{di_2}{dt} \qquad (10\text{-}8)$$

A word about signs is in order. These equations were written for Fig. 10-3, in which winding directions were explicitly shown and could be related to the reference directions of current, voltage, and flux. Suppose one of the winding directions (say on coil 2) had been reversed. A little thought will show that this is equivalent to changing the reference directions of i_2 and v_2. The above equations would become

$$v_1 = L_1 \frac{di_1}{dt} - M \frac{di_2}{dt}$$
$$-v_2 = M \frac{di_1}{dt} - L_2 \frac{di_2}{dt}$$

---

* (continued)
where in the second equation we have used the fact that $\bar{I}_1\bar{I}_2^* = (\bar{I}_2\bar{I}_1^*)^*$. From physical considerations, since neither coil has resistance, the total power (P_1+P_2) must be zero, giving

$$\text{Im}\left[(-k_{12}N_1 + k_{21}N_2)\bar{I}_2\bar{I}_1^*\right] = 0$$

and hence

$$k_{12}N_1 = k_{21}N_2$$

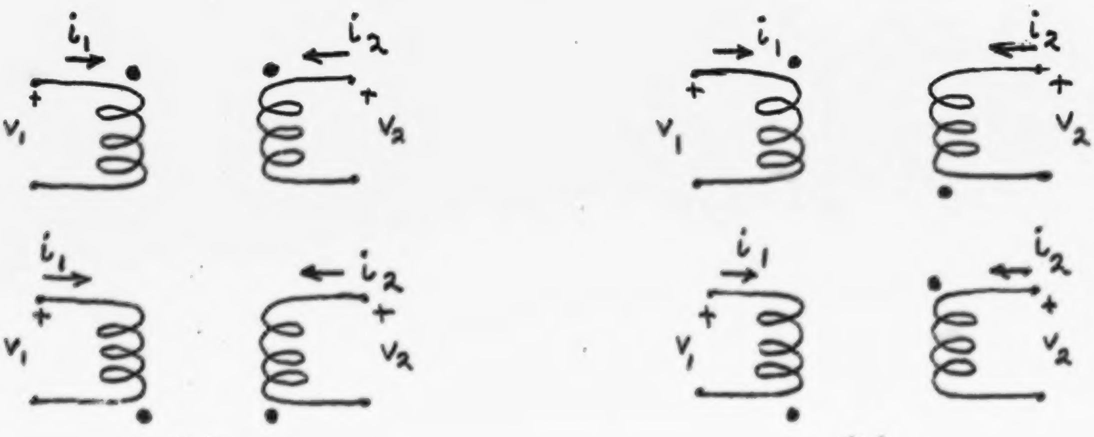
or

$$v_1 = L_1 \frac{di_1}{dt} - M \frac{di_2}{dt}$$

$$v_2 = -M \frac{di_1}{dt} + L_2 \frac{di_2}{dt} \qquad (10\text{-}9)$$

Thus, the sign associated with M becomes negative in both equations. In certain cases, where coils are changing relative position (say one is rotating) it is sometimes convenient to take this effect into account by allowing M itself to become negative, in which case Eqs. (10-8) are perfectly general. We shall always regard M as positive in this text, however.

The question arises as to how to know whether to use + or - signs, when coils are merely shown on a circuit diagram in symbolic form. This is done by the convention of using a pair of dots, as in Fig. 10-4). The two cases shown at (a) will yield Eqs. (10-8), and Eqs. (10-9) apply to the cases shown at (b).



(a) (b)

Figure 10-4.

Finally, observe that these equations and examples have all been given for cases where voltage + references appear at terminals where the current reference arrow enters. A change in any one of these will result in corresponding sign changes in the equations. For example, if the direction of i_2 were reversed in Fig. 10-4a, the equations would be

$$v_1 = L_1 \frac{di_1}{dt} - M \frac{di_2}{dt}$$

$$v_2 = M \frac{di_1}{dt} - L_2 \frac{di_2}{dt}$$

Observe that this is different than the effect of changing the winding direction (because only one reference was changed).

With this information, you should be able to write the equations relating voltages and currents for a pair of mutually coupled coils suitably specified in the manner of Fig. 10-4. Admittedly, it is easy to get confused on signs, especially when the reference directions are not like those in Fig. 10-4. Until you acquire skill in this matter, it is not a bad idea to write the equations first for these reference directions, and then change signs in terms containing any variables which have different references.

10-3. Effect of Resistance

The voltages we have been using are terminal voltages, in the absence of resistance, and are really emf's. For the current reference directions we have been using, the terminal voltage is the emf plus the Ri product. Thus, when coils have respective resistances R_1 and R_2, the complete equations are

$$\begin{aligned} v_1 &= R_1 i_1 + L_1 \frac{di_1}{dt} \pm M \frac{di_2}{dt} \\ v_2 &= \pm M \frac{di_1}{dt} + R_2 i_2 + L_2 \frac{di_2}{dt} \end{aligned} \qquad (10\text{-}10)$$

10-4. Measurement of M

One way to determine M is to connect the two coils in series, as in Fig. 10-5.

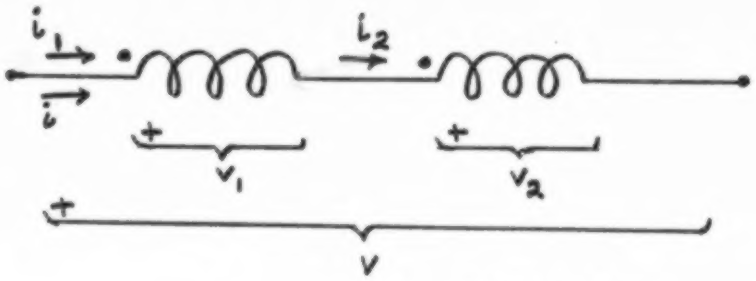


Figure 10-5.

The currents are the same ($i = i_1 = i_2$) and the voltage across the combination is $v = v_1 + v_2$. Thus, Eqs. (10-10) become

$$v_1 = R_1 i + (L_1+M)\frac{di}{dt}$$

$$v_2 = (L_2+M)\frac{di}{dt} + R_2 i$$

and so

$$v = (R_1+R_2)i + (L_1+L_2+2M)\frac{di}{dt}$$

From this equation we reach the conclusion that the inductance of the combination is

$$L_{eq.} = L_1 + L_2 + 2M \qquad (10\text{-}11)$$

Since L_1, L_2, and $L_{eq.}$ can each be measured, M can be found. (What would the last equation be if one of the dots in Fig. 10-5 had been on the other end?)

10-5. Sinusoidal Solutions

If all voltages and currents are sinusoidal, namely

$$\begin{aligned} v_1 &= \sqrt{2}\, V_1 \cos(\omega t + \theta_{v1}) \\ i_1 &= \sqrt{2}\, I_1 \cos(\omega t + \theta_{i1}) \\ v_2 &= \sqrt{2}\, V_2 \cos(\omega t + \theta_{v2}) \\ i_2 &= \sqrt{2}\, I_2 \cos(\omega t + \theta_{i2}) \end{aligned} \qquad (10\text{-}12)$$

equations can be written in terms of the phasor defined from the above, which are

$$\begin{aligned} \bar{V}_1 &= V_1 e^{j\theta_{v1}} & \bar{I}_1 &= I_1 e^{j\theta_{i1}} \\ \bar{V}_2 &= V_2 e^{j\theta_{v2}} & \bar{I}_2 &= I_2 e^{j\theta_{i2}} \end{aligned} \qquad (10\text{-}13)$$

Recalling the principles of sinusoidal circuit analysis, it is recalled that if

$$\bar{I}_1 = I_1 e^{j\theta_{i1}} \text{ is the phasor for } i_1 = \sqrt{2}\, I_1 \cos(\omega t + \theta_{i1})$$

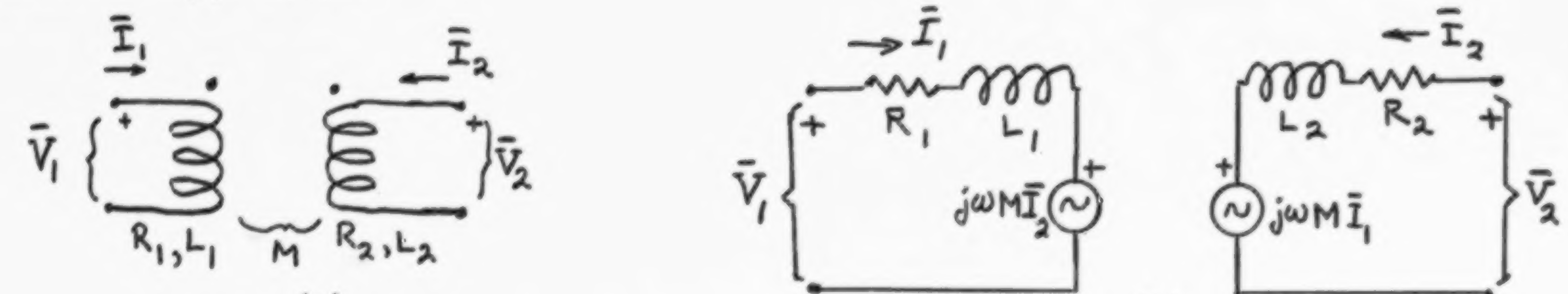
then

$$j\omega\bar{I}_1 \quad \text{is the phasor for} \quad \frac{di_1}{dt}$$

(Refer to the chapter on sinusoids for the derivation of this.) Thus, referring to Eqs. (10-10), in the sinusoidal case only, these can be replaced by

$$\begin{aligned} \bar{V}_1 &= (R_1 + j\omega L_1)\bar{I}_1 \pm j\omega M\bar{I}_2 \\ \bar{V}_2 &= \pm j\omega M\bar{I}_1 + (R_2 + j\omega L_2)\bar{I}_2 \end{aligned} \qquad (10\text{-}14)$$

These equations make it possible to draw an equivalent circuit diagram in which the effect of M is replaced by a pair of dependent voltage sources, as shown in Fig. 10-6b.



(a)

(b)

Figure 10-6.

The original circuit is shown at (a). In Fig. 10-6b there is no mutual coupling; the two parts are entirely separate. But it will be observed that the dependent source in the number (1) circuit depends on $\bar{I}_2$, and vice-versa. If you will write Kirchhoff's equations for Fig. 10-6b, you will obtain the version of Eqs. (10-14) having the + signs. (How would Fig. 10-6b be changed if one of the dots in the original circuit had been at the bottom? Does it make any difference which one?)

This equivalent circuit is not necessary, it provides no information not given by Eqs. (10-14). However, it gives exactly the same information as the equations, and may be easier to remember. Also, when the pair of coils is connected to other circuits, the equivalent circuit can sometimes be a trifle simpler to use. The example given in the next section is a case in point. (See if you can do it without the equivalent circuit.)

10-6. Example of a Solution of a Circuit Problem

Referring to Fig. 10-7a, assume that we are to find the power delivered to the resistor R_L. The equivalent circuit is shown in Fig. 10-7b. (Why is the polarity on the dependent source reversed compared with Fig. 10-6b?)

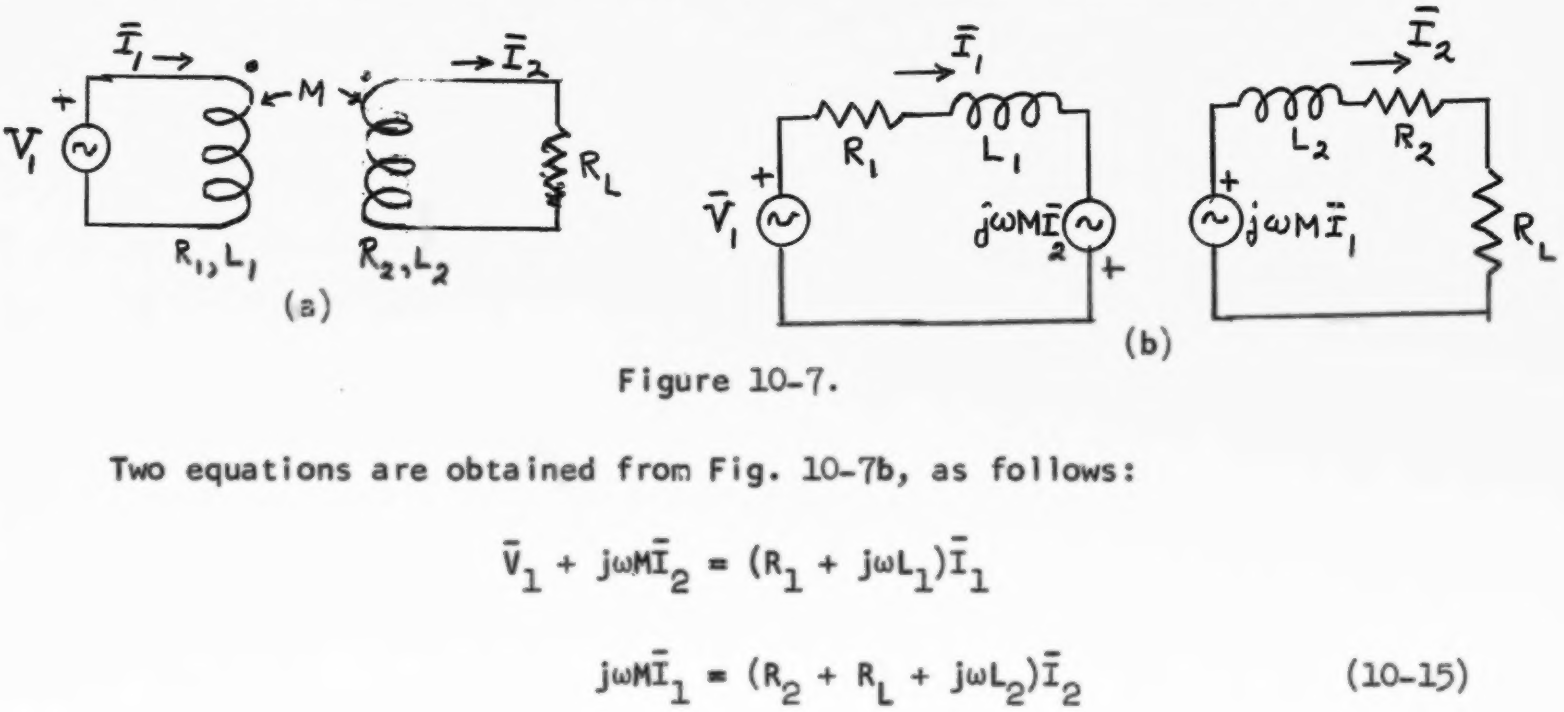


Figure 10-7.

Two equations are obtained from Fig. 10-7b, as follows:

$$\bar{V}_1 + j\omega M\bar{I}_2 = (R_1 + j\omega L_1)\bar{I}_1$$

$$j\omega M\bar{I}_1 = (R_2 + R_L + j\omega L_2)\bar{I}_2 \qquad (10\text{-}15)$$

The second of these gives

$$\bar{I}_1 = \left(\frac{R_2 + R_L + j\omega L_2}{j\omega M}\right)\bar{I}_2$$

$$= \left(\frac{L_2}{M} - j\,\frac{R_2 + R_L}{\omega M}\right)\bar{I}_2$$

This can be substituted into the first of Eqs. (10-15), and after some purely algebraic manipulation, the result is

$$\bar{V}_1 = \frac{1}{M}\left\{R_1L_2+(R_2+R_L)L_1 + \frac{i}{\omega}\left[\omega^2(L_1L_2-M^2)-R_1(R_2+R_L)\right]\right\}\bar{I}_2$$

This of course can be solved for $\bar{I}_2$, giving

$$\bar{I}_2 = \frac{M\bar{V}_1}{R_1L_2 + (R_2+R_L)L_1 + \frac{j}{\omega}\left[\omega^2(L_1L_2-M^2)-R_1(R_2+R_L)\right]}$$

To find the power, we want $|\bar{I}|^2$, which is $M^2|\bar{V}_1|^2$ divided by the square of the absolute value of the denominator. The latter is the sum of the squares of the real and imaginary parts. Thus,

$$P = \frac{M^2|V_1|^2\, R_L}{\left[R_1L_2+(R_2+R_L)L_1\right]^2 + \left[\omega(L_1L_2-M^2) - \frac{R_1(R_2+R_L)}{\omega}\right]^2} \qquad (10\text{-}16)$$

Although this is a long formula, it is simple enough to calculate. Observe that it shows that no power will be transmitted if M = 0, which is an expected result since this is the condition of zero coupling.

10-7. The Iron Core Transformer

The equations we have derived apply to a transformer with an iron core, provided it is valid to assume a linear magnetization curve. However, it is worthwhile to give some special consideration to a simplified version which gives good approximate answers when k is very small, as is usually the case for an iron core. This treatment brings the turns ratio N_1/N_2 into evidence as a fundamental transformer parameter.

Refer to Fig. 10-8, which may be considered the equivalent of Fig. 10-3, except that an iron core is shown explicitly. For simplicity, we shall assume zero resistance in each winding. Dots are shown on the windings, even though they are redundant in view of the fact that winding directions are shown explicitly.

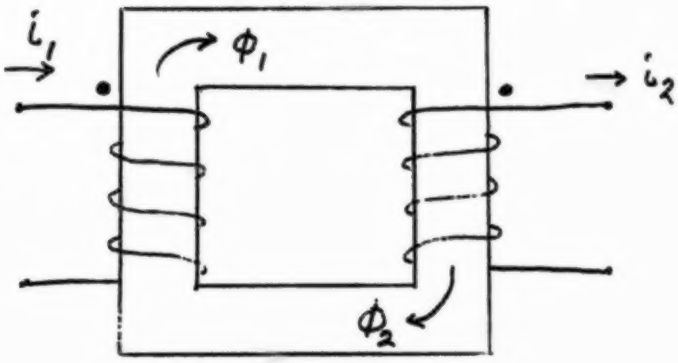


Figure 10-8.

The reference direction of i_2 is reversed compared with Fig. 10-3. This change is not necessary, but is made because the direction in Fig. 10-8 is the one customarily made in the analysis of iron core transformers.

In this case, Eqs. (10-2) and (10-3) become

$$\phi_1 = k_{11}i_1 - k_{12}i_2 \quad , \quad \phi_2 = k_{21}i_1 - k_{22}i_2$$

We now make the crucial assumption that ϕ_1 and ϕ_2 differ by a negligible amount. That is, we assume flux is confined to the core. Then, by equating the above, we get

$$k_{11}i_1 - k_{12}i_2 = k_{21}i_1 - k_{22}i_2$$

or: (10-17)

$$(k_{11}-k_{21})i_1 = (k_{12}-k_{22})i_2$$

This equation must be true for all values of i_1 and i_2. For example, i_1 could be a direct current, and i_2 could be zero; there is no physical reason why this could not be true. Under what condition can the two sides of Eq. (10-17) be equal even for all possible values of i_1 and i_2? Only if each quantity in parentheses is zero, giving

$$\frac{M}{N_2} = k_{21} = k_{11} = \frac{L_1}{N_1} \qquad \text{(a)}$$

(10-18)

$$\frac{M}{N_1} = k_{12} = k_{22} = \frac{L_2}{N_2} \qquad \text{(b)}$$

Referring to Eq. (10-6) we see now that (for the assumption that $\phi_1 = \phi_2$)

$$k = 1$$

and furthermore, from Eq. (10-7),

$$M = \sqrt{L_1L_2} \qquad (10\text{-}19)$$

Now, let us consider ϕ again, which, in view of these results, becomes

$$\begin{aligned}\phi &= k_{11}i_1 - k_{12}i_2 \\ &= \frac{L_1}{N_1} i_1 - \frac{M}{N_1} i_2 \\ &= \frac{L_1}{N_1} i_1 - \frac{\sqrt{L_1L_2}}{N_1} i_2\end{aligned}$$

This gives a relationship between i_2 and i_1 as follows:

$$i_1 = \sqrt{\frac{L_2}{L_1}}\, i_2 + \frac{N_1}{L_1} \phi$$

However, Eq. (10-18b) gives

$$\frac{\sqrt{L_1L_2}}{N_1} = \frac{L_2}{N_2}$$

or

$$\sqrt{\frac{L_2}{L_1}} = \frac{N_2}{N_1} \qquad (10\text{-}20)$$

so that

$$i_1 = \frac{N_2}{N_1} i_2 + \frac{\phi}{k_{11}} \qquad (10\text{-}21)$$

The parameter k_{11} is a function of properties of the magnetic circuit. Specifically, if ℓ is its length, A its cross sectional area, and k_m the relative permeability, solution of the magnetic circuit (considering coil 1 only) gives

$$\phi = \frac{k_m \mu_o A N_1 i_1}{\ell}$$

and so

$$k_{11} = \frac{k_m \mu_o A N_1}{\ell} \qquad (10\text{-}22)$$

If k_{11} is sufficiently large, or Φ is sufficiently small, the Φ/k_{11} term in Eq. (10-21) can be negligible. In that case, we get the approximate equation

$$i_1 = \frac{N_2}{N_1} i_2 \qquad (10\text{-}23)$$

In other words, the current ratio i_1/i_2 is approximately the inverse of the turns ratio.

10-8. The Iron Core Transformer in the Sinusoidal Steady State

We can get a more satisfying answer as to whether Φ/k_{11} is negligible in Eq. (10-21) when all quantities vary sinusoidally. Assume again that all voltages and currents are sinusoidal. Flux will then also be sinusoidal and will be represented by the phasor $\bar{\Phi}$. In Eq. (10-21) we can replace each instantaneous quantity (sinusoidal case only) by the corresponding phasor. Also, we shall use N_1/L_1 in place of $1/k_{11}$, giving

$$\bar{I}_1 = \frac{N_2}{N_1} \bar{I}_2 + \frac{N_1}{L_1} \bar{\Phi}$$

However, we also have

$$\bar{V}_1 = j\omega N_1 \bar{\Phi} \qquad (10\text{-}24)$$

which combines with the above to give

$$\bar{I}_1 = \frac{N_2}{N_1} \bar{I}_2 + \frac{\bar{V}_1}{j\omega L_1} \qquad (10\text{-}25)$$

Thus, in this case we see that

$$\frac{\bar{I}_1}{\bar{I}_2} = \frac{N_2}{N_1} \qquad (10\text{-}26)$$

if

$$\frac{\bar{V}_1}{j\omega L_1}$$

is small. This quantity, is called the magnetizing current. It is the current in winding 1 when it is connected to a source of voltage $\bar{V}_1$ and when winding 2 is an open circuit.

Corresponding to Eq. (10-24), we also have

$$\bar{V}_2 = j\omega N_2 \bar{\Phi}$$

and so we see that under the assumptions being made (zero resistance and no leakage flux)

$$\frac{\bar{V}_1}{\bar{V}_2} = \frac{N_1}{N_2} \tag{10-27}$$

If winding resistance is introduced, Eq. (10-27) becomes an approximation.

Equation (10-24) provides the point of departure for the last consideration. Let us write it as an absolute value relation (using symbols without bars for absolute values). Thus, we can write

$$\Phi = \frac{V_1}{\omega N_1} \tag{10-28}$$

This is crucial because there is a limit to the value that Φ can attain for approximat linear operation in a transformer with iron core.

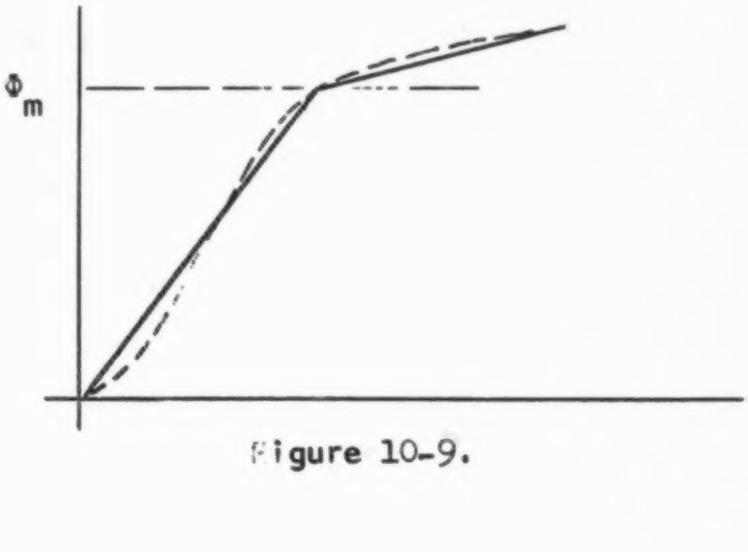


Figure 10-9.

In Fig. 10-9 is shown a linearized approximation of the magnetization curve of the magnetic circuit. It will not be permissible to exceed a certain value of flux, say Φ_m. This means that the Φ in Eq. (10-28) must be less than $\Phi_m/\sqrt{2}$ (Why?). In terms of voltage, frequency, and number of turns, this gives

$$\frac{V_1}{\omega N_1} < \frac{\Phi_m}{\sqrt{2}}$$

This relationship has useful and important implications concerning the use of a transformer on other than the voltage or frequency for which it was designed. Suppose it was designed so that at rated voltage and frequency, the above relationship was just barely satisfied. If the voltage is raised, the limit will be exceeded. It will also be exceeded if the frequency is decreased, while the voltage remains unchanged. The only way to operate a transformer at reduced frequency is to reduce the voltage in proportion.

Equation (10-28) also shows why it is expensive to make "audio transformers" that will operate well at low frequencies. They must be provided with massive magnetic circuits that will permit the large Φ demanded by Eq. (10-28) when ω is small. Parenthetically, it might be suspected that good performance at low frequencies would be provided by making N_1 large. This would be true, except that such transformers must also operate at high frequencies, and when N_1 is made too large interwinding capacitance becomes detrimental at high frequencies. Design of transformers that will operate over a wide range of frequencies is not a job for the uninitiated.